processes

MDPI

Article

Multi Time Scale Continuous Wavelet Transform Ensemble Learning for Acoustic Emission-Based Tensile–Shear Fracture Mode Identification

Shuo Yang [1], Qian Jiang [2], Chengxu Yang [2], Zilong Chen [2], Xiangzheng Luo [3], Chao Liu [3,*] and Lei Wang [2,*]

[1] School of Civil Engineering, Xuzhou University of Technology, Xuzhou 221018, China; yangshuo@xzit.edu.cn
[2] School of Computer Science and Technology/School of Artificial Intelligence, China University of Mining and Technology, Xuzhou 221116, China; qianjiang@cumt.edu.cn (Q.J.); ts25170031a31@cumt.edu.cn (C.Y.); chenzilong@cumt.edu.cn (Z.C.)
[3] School of Mechanics and Civil Engineering, China University of Mining and Technology, Xuzhou 221116, China; ts26030183p31@cumt.edu.cn
* Correspondence: liuc_rm@cumt.edu.cn (C.L.); leiwang@cumt.edu.cn (L.W.)

Abstract

Accurate identification of tensile and shear acoustic emission (AE) modes is important for revealing rock damage–fracture mechanisms and early warning of dynamic hazards. The conventional RA–AF empirical criterion relies on manually defined boundaries and cannot stably separate overlapping point clouds. CWT time–frequency images with deep learning offer an end-to-end alternative, but extraction-window duration is usually set empirically and its independent effect remains unclear. This study proposes a multi-time-scale AE framework (MTS-AE) that combines signal-level multi-scale detection, event-level strictly paired extraction and hierarchically locked partitioning, and representation-level unified CWT parameters, forming a model matrix of eight deep-learning architectures and seven window durations (0.5–10 ms). Odds ratio analysis reveals architecture-dependent class-conditional window preferences: ResNet18 prefers 3 ms for tensile-associated records (OR = 3.86) and 10 ms for shear-associated records (OR = 2.73). Same-architecture cross-window decision-level integration yields positive net gains for all eight architectures, improving ResNet18 from 98.86% to 99.75%; a seven-window ResNet18 ensemble (99.75%) slightly exceeds the full 56-model ensemble under the present evaluation protocol at one-eighth the number of ensemble members. These findings demonstrate the value of complementary temporal representations for AE classification and provide guidance for extraction-window design and multi-window ensemble learning.

Keywords: acoustic emission; multi-time-scale analysis; continuous wavelet transform; fracture-mode identification; decision-level ensemble

Check for updates

Academic Editor: Dicho Stratiev

Received: 9 October 2026
Revised: 26 October 2026
Accepted: 29 October 2026
Published: 3 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Acoustic emission (AE) is a physical phenomenon in which strain energy is instantaneously released and propagates outward in the form of transient elastic waves as internal microcracks initiate, propagate, and eventually coalesce within a material under loading. Because this process carries real-time information concerning crack location, material strength, and fracture mechanisms, AE has become an important nondestructive monitoring technique for investigating damage evolution in brittle materials such as rock and concrete [1–3]. In rock fracture, AE signals associated with tensile-dominated Mode I cracking and shear-dominated Mode II cracking may exhibit different characteristics in terms of

energy, frequency, and duration. Dynamic variations in the relative proportions of these two crack types can therefore serve as important precursor indicators of the critical transition from distributed microscopic damage to macroscopic instability [4,5]. Accordingly, the analysis of AE signal characteristics associated with different fracture modes is important for investigating rock-fracture mechanisms and improving the early warning of engineering hazards associated with rockbursts, fault instability, and deep mining [6,7]. However, AE signals are non-stationary, broadband, and characterized by low signal-to-noise ratios. Most individual events release their energy within several milliseconds, span frequency ranges of several hundred kilohertz, and are susceptible to environmental noise. These characteristics have long posed substantial challenges to the automatic and reliable identification of fracture modes.

Recent advances in rock fracture mechanics have further demonstrated that tensile and shear cracking are not mutually exclusive during progressive rock failure. Wu et al. [8] investigated mixed-mode I/II fracture in quartz-diorite using digital image correlation and fracture-toughness analysis, revealing that crack propagation may evolve from an initially mixed-mode condition toward Mode I failure. Khadivi et al. [9] combined acoustic emission monitoring with high-speed imaging to investigate the fracture processes of different rock types under Brazilian splitting, showing that tensile microcracking generally dominates the initial fracture stage, while shear microcracking may subsequently develop. More recently, Heng et al. [10] investigated the continuous–discontinuous failure characteristics of granite under Brazilian splitting and demonstrated that loading conditions influence crack initiation and fracture-process-zone evolution. Under compressive–shear loading, Zhou et al. [11] employed digital image correlation to distinguish tensile, shear, and mixed-mode cracks, demonstrating that pre-existing flaw geometry significantly affects crack propagation and coalescence. Collectively, these studies indicate that rock fracture is governed by interacting tensile and shear mechanisms, and that the dominant failure mode depends on the loading conditions and fracture evolution process. These findings provide a rock-mechanics basis for understanding the complexity of fracture-mode identification using AE signals.

Early fracture-mode identification methods relied primarily on empirical criteria derived from time-domain parameters of AE waveforms. Among these, the rise-time-to-amplitude ratio and average-frequency (RA–AF) criterion has been widely used [12]. Based on concrete experiments, Ohno and Ohtsu [13] distinguished tensile and shear cracks using a dividing line in the RA–AF coordinate system. Aggelis [14] further observed that AE-event characteristics tended to shift toward shear-dominated behavior as concrete approached failure. In rock mechanics, Du et al. [4] systematically compared the evolution of the proportions of the two microcrack types under Brazilian splitting, three-point bending, direct shear, and uniaxial compression. Zhang and Deng [15] proposed a new method for determining the classification boundary based on waveform-frequency characteristics, whereas Dong et al. [16] investigated the uncertainty in RA–AF classification results caused by variations in AE source scale. These methods have clear physical interpretations and are computationally efficient. However, the slope of the classification boundary is generally determined empirically. When the number of events is large and the underlying fracture mechanisms are complex, the AE data associated with the two fracture categories tend to overlap near the dividing region, making it difficult for a single linear boundary to achieve stable separation.

To reduce the dependence of parameter-based methods on manually specified thresholds, machine-learning techniques were subsequently introduced. Das et al. [17] combined unsupervised clustering with a support vector machine (SVM) to classify fracture modes without requiring manual annotation. Ju et al. [18] systematically compared the performance of several clustering algorithms for crack classification. For rock materials,

Qin et al. [19] extracted time–frequency features from the Wigner–Ville distribution of limestone AE signals and combined them with an SVM for classification. Wang et al. [20] developed a particle-swarm-optimization Gaussian-mixture-model support-vector-machine framework, abbreviated as PSO-GMM-SVM, to classify RA and AF parameters derived from direct shear tests on granite. Although these methods partially overcome the limitations of empirical criteria, their discriminative capability remains strongly dependent on manually designed feature sets, and the detailed time–frequency information embedded in the original waveforms is not fully utilized.

To exploit waveform information more comprehensively, end-to-end methods combining time–frequency representations with deep learning have emerged in recent years. Their basic principle is to transform one-dimensional transient waveforms into two-dimensional time–frequency images using methods such as the short-time Fourier transform and continuous wavelet transform (CWT) [21], after which convolutional neural networks (CNNs) or related models are employed to automatically extract discriminative features. Sun et al. [22] extracted features using the short-time Fourier transform and Mel spectrograms and developed a Mel + VGG model for the real-time identification of critical signals. Using data obtained from Brazilian splitting and direct shear tests on red sandstone, Ren et al. [23] compared several CNNs that used waveform plots and wavelet-packet time–frequency spectra as inputs for tensile/shear classification. For composite materials, Sikdar et al. [24] transformed AE waveforms recorded from a composite panel into CWT scalograms and employed a CNN to classify damage-source regions. Li et al. [25] combined synchrosqueezed wavelet transforms with a multi-branch CNN to classify waveforms associated with rail cracks. These studies demonstrate that time–frequency images can effectively preserve the non-stationary and multiscale structures of AE signals. Among the available methods, CWT has been widely adopted because of its strong time–frequency localization capability.

Despite the increasing use of time–frequency representations for AE classification, the selection of event-extraction window duration remains an important design choice that has not always been systematically examined. Window duration determines the temporal coverage of the extracted AE signal and influences the transient and decay components retained in the CWT representation, even when the mother wavelet, frequency axis, and image resolution are held constant. Related research has highlighted the importance of temporal information selection. For example, Song et al. [26] investigated the influence of AE waveform sequence length on the identification of sandstone fracture stages under true triaxial loading and found that both excessively long and excessively short sequences could reduce classification accuracy. However, their sequence length refers to the number of AE waveforms combined rather than the extraction duration of an individual AE event. Thus, the influence of systematically varying the extraction-window duration for the same AE event remains insufficiently characterized in the studies reviewed here. Moreover, the effects of window duration may be confounded by variations in data partitioning, preprocessing parameters, and event detection. Consequently, a controlled analysis using consistently paired AE events and unified preprocessing conditions is needed to better isolate the contribution of extraction-window duration.

To address these issues, this study treats event-extraction window duration as a key representation-design parameter and systematically investigates its influence on AE classification under controlled experimental conditions. The investigation is organized into three complementary aspects: methodological framework, model-based window-preference analysis, and ensemble implementation strategy.

At the methodological level, a multi-time-scale acoustic emission analysis method, termed MTS-AE, is proposed. The framework combines consistent event-detection anchors, paired waveform extraction across seven window durations, and unified CWT preprocessing

parameters. By maintaining consistent data organization and representation settings across the investigated windows, the proposed method reduces potential confounding effects associated with event detection, data partitioning, and preprocessing, thereby providing a controlled framework for evaluating window-duration effects.

At the model-analysis level, a model matrix comprising eight representative deep-learning architectures and seven candidate window durations is constructed. An odds-ratio (OR) metric is employed to characterize the relative class-conditional window preferences derived from model predictions on the balanced CWT dataset. The resulting preferences vary across architectures, indicating that the discriminative contribution of temporal coverage is model-dependent. In the present binary formulation, the class-specific odds ratios are reciprocal; therefore, the non-overlap between their extrema should not be interpreted as an independent physical invariant. For ResNet18, the strongest relative preferences occur at 3 ms for tensile-associated records and 10 ms for shear-associated records within the investigated window set. These model-derived window preferences should not be equated with the intrinsic physical durations of individual fracture events.

At the implementation-strategy level, a same-architecture, cross-window-duration decision-level ensemble scheme is developed to exploit complementary information from different temporal representations. Comparative experiments demonstrate positive net classification gains for all eight evaluated architectures under the adopted evaluation protocol. The results indicate that window-duration diversity can provide an effective ensemble dimension alongside architectural diversity. Based on these findings, a potential engineering implementation strategy using a single high-performing architecture with parallel multi-window inference is proposed, although its practical real-time performance requires further field validation.

2. MTS-AE Method

2.1. Overall Framework

To rigorously isolate the effect of window duration from confounding factors, including perturbations in data partitioning, preprocessing inconsistencies, and event-detection drift, this study proposes a multi-time-scale acoustic emission analysis method, termed MTS-AE. As illustrated in Figure 1, the method comprises three progressive processing levels: the signal layer, the event layer, and the representation layer. These levels establish an end-to-end mapping from raw waveforms to CWT time–frequency-image classification. A cross-window-duration decision-level ensemble module is further incorporated to form a complete methodological framework.

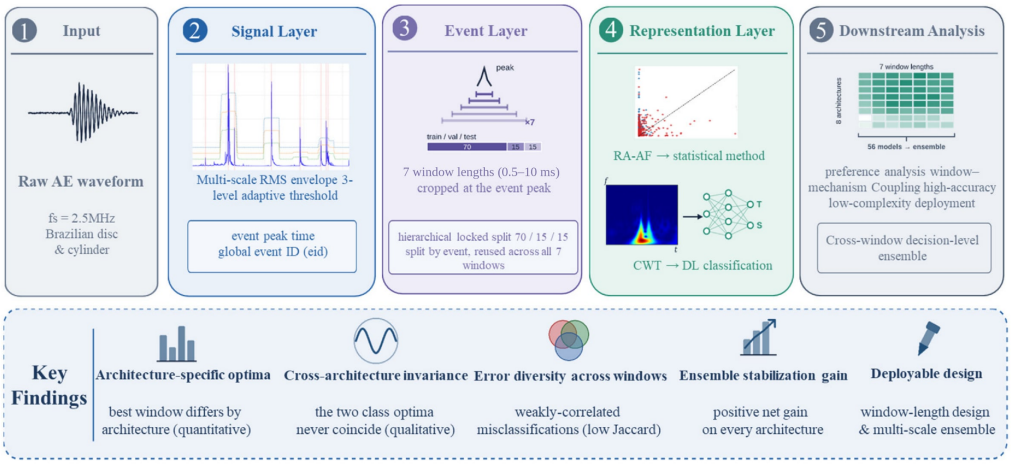


Figure 1. Overall framework of the MTS-AE method.

As shown in Figure 1, the MTS-AE workflow sequentially performs multi-scale event detection at the signal layer, strictly paired multi-time-scale extraction at the event layer, and cross-window-comparable CWT time–frequency mapping at the representation layer. The procedure ultimately generates seven mutually corresponding CWT datasets, each associated with a specific window duration, thereby providing standardized inputs for subsequent classification and decision-level ensemble learning.

The central methodological principle of MTS-AE is to treat event-extraction window duration as a controlled representation variable rather than an independently selected preprocessing parameter. Unlike conventional multi-window approaches that may introduce inconsistencies in event extraction, data partitioning, and signal representation, MTS-AE establishes a unified analytical framework in which the same physical AE events are examined across different temporal scales. Specifically, the signal layer provides common event-detection anchors, the event layer maintains one-to-one correspondence and consistent data partitions across window durations, and the representation layer applies unified CWT settings to construct comparable time–frequency representations. By controlling these factors, the proposed framework enables a systematic assessment of the influence of temporal-scale selection on fracture-mode identification while reducing interference from unrelated processing variations. Furthermore, the correspondence among window-specific representations provides the methodological basis for exploiting complementary temporal information through decision-level ensemble learning.

2.2. Multi-Scale Event Detection

AE waveform sequences generally span long time intervals and exhibit non-stationary background noise. A threshold defined at a single time scale therefore cannot adequately accommodate events with different amplitudes and durations. To address this issue, the proposed method employs an event-detection strategy combining multi-scale root-mean-square (RMS) envelopes with three-level adaptive thresholds. Let the raw waveform be denoted by $x(\mathrm{t})$. The local RMS envelope at time scale τ is defined as:

$$E_\tau(t) = \sqrt{\frac{1}{N_\tau} \sum_{k=t-\frac{N_\tau}{2}}^{t+\frac{N_\tau}{2}} x(k)^2}, N_\tau = \lfloor \tau \cdot f_s \rfloor \tag{1}$$

where $\tau \in \{0.1, 0.2, 0.5, 1.0\}$ ms represents the four time scales, and f_s denotes the sampling frequency, as detailed in Section 3.1, and $\lfloor\ \rfloor$ denotes the floor function.

The short-scale envelopes at 0.1 and 0.2 ms preserve rising-edge information associated with sharp transients, whereas the long-scale envelopes at 0.5 and 1.0 ms suppress short-term fluctuations while retaining the overall energy structure of an event. Their parallel use avoids introducing a bias toward any single time window.

Event detection is conducted using a three-level adaptive threshold:

$$T_{\mathrm{local}}^{(k)}(t) = \mu_{\mathrm{local}}(t) + k \cdot \sigma_{\mathrm{local}}(t), k \in \{3.5,\ 2.5,\ 1.5\} \tag{2}$$

where $k = 3.5$, 2.5, and 1.5 correspond to the major, medium, and minor intensity levels, respectively. The local mean μ_{local} and local standard deviation σ_{local} are estimated within a sliding window of W = 20 ms.

To prevent false detections caused by excessively low thresholds in low-noise regions, an additional global safeguard is imposed:

$$T(t) = \max\left[T_{\mathrm{local}}^{(k)}(t),\ \mu_{\mathrm{global}} + k_{\min}\sigma_{\mathrm{global}}\right] \tag{3}$$

where T(t) is the globally safeguarded threshold, μ_{global} and σ_{global} are the global mean and standard deviation, respectively, and k_{min} is the global threshold coefficient. In this study, $k_{min} = 1.5$, consistent with the coefficient used for the minor level.

Detection results from the three threshold levels are merged according to the priority order: major > medium > minor. Adjacent detections are deduplicated by retaining the event with the largest amplitude. The minimum inter-event interval is set to 0.5 ms, the minimum peak amplitude is 5×10^{-5} V, and the maximum allowable proportion of zero-valued samples is 0.90.

2.3. Strictly Paired Multi-Time-Scale Extraction

Using the events detected at the signal layer as anchors, the event layer maps each event onto multiple time scales so that waveform segments corresponding to the same physical event can be extracted under different window durations.

2.3.1. Multi-Window Event-Waveform Sets

For each window duration $L \in \mathcal{L} = \{0.5, 1, 2, 3, 4, 5, 10\}$ ms, a waveform segment of duration L is extracted around the event peak. The resulting waveform set is defined as:

$$\mathcal{W}_L = \left\{ \left(w_i^{(L)}, y_i, \mathrm{eid}_i \right) \right\}_{i=1}^{N} \tag{4}$$

where $w_i^{(L)}$ denotes the waveform segment of the i-th event extracted using window duration L, $y_i \in \{0,1\}$ denotes the class label, with 0 representing AE events associated with tensile-dominated failure and 1 representing AE events associated with shear-dominated failure, and eid_i is the globally unique event identifier. N denotes the total number of events.

The peak time of each event is held constant across all window durations. Thus, waveform segments obtained under different windows are aligned to the same physical instant. The selection of window durations was guided by the multiscale temporal characteristics of AE signals and the need to balance time-scale coverage with computational efficiency. Accordingly, seven representative window durations, L = {0.5, 1, 2, 3, 4, 5, 10} ms, were selected to investigate the influence of temporal coverage on fracture-mode identification. The relatively dense sampling between 0.5 and 5 ms was intended to characterize the sensitivity of classification performance to short- and intermediate-duration signal components, particularly the transient response and early decay phase. These durations provide different levels of temporal coverage for examining localized time–frequency characteristics associated with AE events.

A longer window of 10 ms was additionally included to investigate the potential contribution of extended decay components, which may contain discriminative information associated with shear-dominated fracture processes. Intermediate durations between 5 and 10 ms were not included to limit the computational cost while retaining representative short-, intermediate-, and long-duration configurations. Furthermore, excessively long extraction windows may increase the possibility of incorporating signals from neighboring events. Thus, the selected durations were intended to provide a representative rather than exhaustive exploration of the window-duration parameter space. The optimal durations identified in this study should therefore be interpreted within the investigated candidate set.

2.3.2. Hierarchically Locked Partitioning

To ensure that cross-window performance differences can be attributed to window duration rather than to data-partition perturbations, an event-ID-based hierarchically locked partitioning strategy is employed.

Under a fixed random seed, all event IDs are partitioned only once:

$$\mathcal{I} = \mathcal{I}_{\text{train}} \sqcup \mathcal{I}_{\text{val}} \sqcup \mathcal{I}_{\text{test}},\ |\mathcal{I}_{\text{train}}|:|\mathcal{I}_{\text{val}}|:|\mathcal{I}_{\text{test}}| = 7{:}1.5{:}1.5 \tag{5}$$

where $\mathcal{I}$ denotes the complete event-ID set and $\sqcup$ denotes a disjoint union.

The resulting partition is then applied unchanged to all seven window durations. For example, when event eid_i belongs to the test subset of $\mathcal{W}_{0.5}$, it must also belong to the test subset of $\mathcal{W}_{10}$. This mechanism freezes, across all window durations, the random assignment of each event to the training, validation, or test subset. Consequently, models evaluated under the seven window durations encounter exactly the same event sets, thereby eliminating one of the largest potential sources of bias in cross-window comparisons.

To ensure comparability across window durations, a fixed-seed partitioning strategy was applied consistently to the CWT records used for the seven temporal windows. The same record-level subset assignments were maintained across windows, thereby controlling variation associated with data partitioning in the cross-window comparisons.

2.4. Cross-Window-Comparable CWT Time-Frequency Mapping

Each event-waveform segment is first preprocessed by removing its direct-current component, applying a 10–1000 kHz bandpass filter, and normalizing its amplitude to the interval $[-1, 1]$. A continuous wavelet transform is subsequently performed using the complex Morlet wavelet cmor1.5-1.0:

$$W(a,b) = \int_{-\infty}^{+\infty} x(t) \frac{1}{\sqrt{a}} \psi^* \left(\frac{t-b}{a} \right) dt \tag{6}$$

where a is the scale factor, b is the translation parameter, and $\psi*$ denotes the complex conjugate of the mother wavelet. Scale and frequency are related by $f = f_c \cdot f_s / a$. where fc is the center frequency of the mother wavelet. The frequency-analysis range is defined as $f \in [10, 1000]$ kHz, which covers the dominant frequency range of typical AE events in rock. The frequency dimension is uniformly sampled at 256 points. For each event, the CWT magnitude matrix $|W(a,b)|$ is mapped onto a 256 × 256 two-dimensional grid. Following channel-wise normalization, the matrix is rendered as an RGB pseudocolor image using the jet colormap:

$$\mathcal{D}_L = \left\{ \left(X_i^{(L)}, y_i, \text{eid}_i \right) \right\}_{i=1}^{N}, X_i^{(L)} \in \mathbb{R}^{3 \times 256 \times 256} \tag{7}$$

To eliminate preprocessing inconsistencies across window durations, all seven windows share an identical set of preprocessing parameters. Specifically, the following parameters are locked across window durations: CWT mother wavelet: cmor1.5-1.0; frequency axis: 10–1000 kHz with 256 sampling points; image resolution: 256 × 256; colormap: jet.

Figure 2 presents the CWT representations of a representative tensile event, tensile_00747, and a representative shear event, shear_00034, under the seven window durations. Both events were correctly classified by ResNet18 under all seven durations with confidence scores of at least 0.95 and can therefore be regarded as clean representative samples of their respective classes.

Distinct time–frequency structures can be observed between the two fracture modes. Under the short 0.5 ms window, the low-frequency energy of the tensile event exhibits a broadly dispersed, cluster-like distribution, with wide frequency-band coverage and irregular boundaries. In contrast, the low-frequency energy of the shear event is highly concentrated within a narrower frequency band and forms a relatively symmetric and rounded cluster.

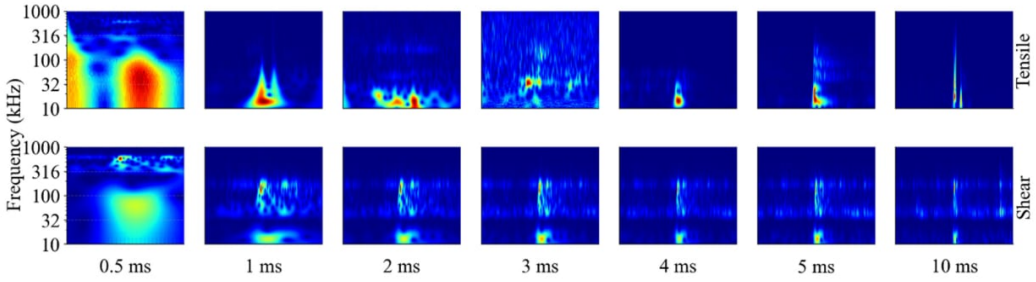


Figure 2. CWT representations of the same tensile event, tensile_00747, and the same shear event, shear_00034, under seven window durations. The horizontal axis represents peak-centered time in milliseconds, and the vertical axis represents logarithmic frequency in kilohertz.The color scale represents the normalized CWT magnitude (energy), with dark blue indicating lower values and red indicating higher values.

As the window duration increases, the tensile event exhibits isolated, spike-like energy concentrations in the high-frequency region, appearing as point-like or short-band structures in the time–frequency plane. The shear event, by contrast, exhibits multiple approximately parallel stripes extending obliquely toward higher frequencies. These structures reflect the characteristic dispersion and sustained oscillations associated with shear fracture.

2.5. Cross-Window-Duration Decision-Level Ensemble

Sections 2.3 and 2.4 establish a one-to-one correspondence among samples in the seven window-specific datasets at both the data-organization and representation levels. On this basis, a same-architecture, cross-window-duration decision-level ensemble scheme is defined. Let $f_{a,L}$ denote the model trained using architecture a and window duration $L \in \mathcal{L}$, and let $p_{a,L}(X)$ denote the softmax probability vector produced by the model for sample X. Four decision-level fusion strategies are compared. Equal-weight probability averaging is adopted as the default strategy:

$$p_a^{(\mathrm{eq})}(X) = \frac{1}{|\mathcal{L}|} \sum_{L \in \mathcal{L}} p_{a,L}(X), \hat{y}^{(\mathrm{eq})}(X) = \arg\, max p_a^{(eq)}(X) \tag{8}$$

Three alternative strategies are evaluated for comparison, with the empirical results presented in Section 4.3.2. The first strategy weights each model according to its validation-set accuracy:

$$p_a^{(\mathrm{val})}(X) = \sum_{L \in \mathcal{L}} \frac{\mathrm{Acc}_{a,L}^{\mathrm{val}}}{\sum_{L' \in \mathcal{L}} \mathrm{Acc}_{a,L'}^{\mathrm{val}}}, p_{a,L}(X) \tag{9}$$

The second strategy applies sample-wise confidence weighting based on the maximum predicted class probability:

$$p_a^{(\mathrm{conf})}(X) = \sum_{L \in \mathcal{L}} \frac{\max\limits_c p_{a,L,c}(X)}{\sum_{L' \in \mathcal{L}} \max\limits_c p_{a,L',c}(X)}, p_{a,L}(X) \tag{10}$$

The third strategy uses hard voting:

$$\hat{y}_a^{(\mathrm{hard})}(X) = \mathrm{mode}\left\{\arg \max_c p_{a,L,c}(X) : L \in \mathcal{L}\right\} \tag{11}$$

Equation (9) corresponds to validation-accuracy weighting, Equation (10) implements sample-wise confidence weighting based on the maximum class probability, and Equation (11) implements hard voting. Equal-weight averaging does not require additional validation-dependent weighting or introduce extra hyperparameters. It is also consistent with the methodological premise that window duration should be treated as a design variable of

equal importance to model architecture. Equal-weight averaging is therefore adopted as the default ensemble strategy throughout this study.

The complete MTS-AE workflow integrates multi-scale event detection, strictly paired multi-window extraction, unified CWT representation, and model training. For each architecture, seven independently trained window-specific models generate prediction probabilities, which are aligned by event identifiers and combined using equal-weight averaging. Detailed experimental settings are provided in Section 3.4.

3. Experimental Design

3.1. Specimens and Acoustic Emission Data Acquisition

Nine Brazilian disk splitting tests and four uniaxial compression tests were conducted to obtain acoustic emission (AE) signals under two different loading conditions. Brazilian splitting is generally associated with tensile-dominated failure, whereas uniaxial compression may involve tensile, shear, and mixed-mode cracking, even when the macroscopic failure is shear-dominated. For the present binary classification task, AE events recorded during Brazilian splitting were assigned to the tensile-associated category, while those recorded during uniaxial compression were assigned to the shear-associated category. These labels were determined by the corresponding experimental loading conditions rather than by independent verification of the microscopic fracture mechanism of each AE event. Consequently, the classification results may partly reflect differences between the two experimental conditions and should not be interpreted as direct identification of the fracture mechanism of every individual event. The numbers of original AE events obtained under the two loading conditions, before oversampling and data augmentation, are summarized in Table 1. These counts refer to the original detected events before class balancing and oversampling.

Table 1. Distribution of original AE events among the specimens under different loading conditions.

Specimen ID	Loading Condition	Dimensions (mm)	Number of Original AE Events	Assigned Category
B01–B09	Brazilian splitting	Φ50 × 25	262	Tensile-dominated
U01–U04	Uniaxial compression	Φ50 × 100	2619	Shear-dominated
Total	13 tests	—	2881	—

The experimental data were acquired using a servo-controlled rock-mechanics testing system (Jinan Sanqin Testing Technology Co., Ltd., Jinan, China). As shown in Figure 3, the system has a maximum loading capacity of 250 kN and supports both force-controlled and displacement-controlled loading. It can be used to conduct various mechanical tests, including Brazilian splitting and uniaxial compression tests. All specimens were prepared from the same batch of marble to minimize the influence of material heterogeneity on the experimental results. The specimens comprised standard cylinders with a diameter of 50 mm and a height of 100 mm, as well as disks with a diameter of 50 mm and a thickness of 25 mm. Loading was applied by the servo-hydraulic testing machine under displacement control. Each test was performed on a separate specimen, with no specimen reused across the Brazilian splitting and uniaxial compression experiments.

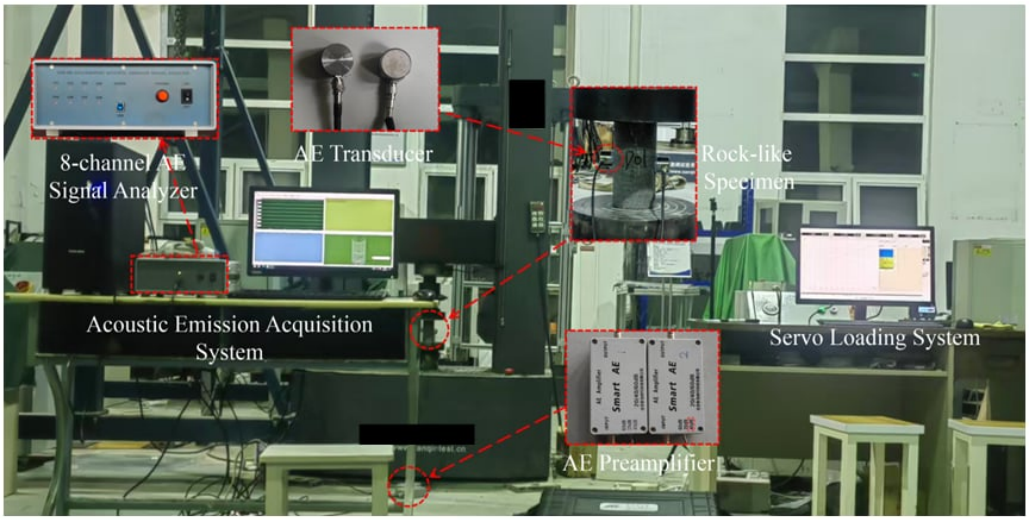


Figure 3. Experimental setup.

AE signals generated during rock failure were acquired using a DS5-8A/B full-waveform acoustic emission signal analyzer(Beijing Softland Times Technology Co., Ltd., Beijing, China). The system has eight independent acquisition channels and operates at a sampling frequency of 2.5 MHz, corresponding to a temporal resolution of $\Delta t = 0.4$ μs. To preserve the complete frequency-domain information of the raw signals, both the hardware analog filter and the digital filter were set to bypass mode. Continuous full-waveform data were recorded over the entire available frequency range without interruption. No amplitude-trigger threshold was applied at the acquisition-device level. Instead, all AE events analyzed in this study were detected and extracted offline from the continuously recorded waveforms using the software-based procedure described in Section 2.2. The maximum duration of a single continuous acquisition was set to the device limit of 10,000 s, although the actual duration of each experiment was substantially shorter.

3.2. Candidate Architectures

To ensure that the generalizability of the multi-window comparison was not restricted by a narrow selection of model architectures, eight representative deep-learning architectures from three model families were implemented and trained under identical experimental conditions, as summarized in Table 2. The KAN blocks replace fixed activation functions with learnable B-spline functions. Their hierarchical formulation is given by $h^{(l+1)} = \Phi^{(l)}(h^{(l)})$, $\Phi_{ij}^{(l)}(x) = \sum_k c_{ijk}^{(l)} B_k(x)$, where B_k denotes a K-th-order B-spline basis function. The state-update equations of the selective state-space model used in Mamba2 are $h_t = A(u_t)h_{t-1} + B(u_t)u_t$, $y_t = C(u_t)h_t$, where A, B, C are input-dependent selective matrices. Bi-Mamba2_row extends Mamba2_row by introducing bidirectional scanning. Unlike convolutional architectures, which explicitly capture local spatial relationships through two-dimensional kernels, the Mamba-based models employ selective state-space mechanisms to model dependencies along sequentially organized features. In the present implementation, the two-dimensional CWT representations are converted into one-dimensional sequences through row-wise scanning, allowing sequential dependency modeling while not explicitly preserving all spatial neighborhood relationships along both the temporal and frequency dimensions. All models have three input channels and two output classes.

Table 2. Overview of the candidate architectures.

Family	Architecture	Number of Parameters	Principal Modeling Assumption
CNN	ResNet18, DenseNet121, EfficientNet-B0, and MobileNetV2	3.5–11.7 M	Local convolutional kernels capture localized time–frequency textures and multiscale structures
KAN	UKAN and KAN_Hybrid_Net comprising ResNet18 with a KAN head	0.19 M and 11.83 M	Learnable B-spline functions refine the decision boundary
Mamba	Mamba2_row and Bi-Mamba2_row using selective state-space modeling with two-dimensional-to-one-dimensional row-wise scanning	0.18 M and 0.28 M	Selective sequence-dependency modeling with linear computational complexity

Their embedding dimensions and network depths were configured according to the corresponding original publications or official implementations [27–34].

3.3. Conventional RA-AF Baseline

To provide a conventional baseline for evaluating the CWT representations, two time-domain parameters derived from the classical method of Ohno and Ohtsu [13] were simultaneously extracted from each event-waveform segment: the RA value and average frequency. They are defined as RA = t_{rise}/A (μs/mV), and AF = N_c/τ_{dur} (kHz), where t_{rise} is the waveform rise time, A is the peak amplitude, N_c is the number of zero crossings during the event, and τ_{dur} is the event duration.

In the event dataset used in this study, tensile events are concentrated in the low-RA and high-AF region, whereas shear events exhibit a more dispersed distribution. However, when plotted on logarithmic coordinates, the two event clouds overlap substantially in their central regions, preventing a single linear boundary from consistently separating the two fracture modes. The substantial overlap between the two categories suggests that a simple linear RA–AF boundary may have limited discriminative capability for the present dataset, motivating the use of richer CWT time–frequency representations. Their detailed distributions are presented in Figure 4.

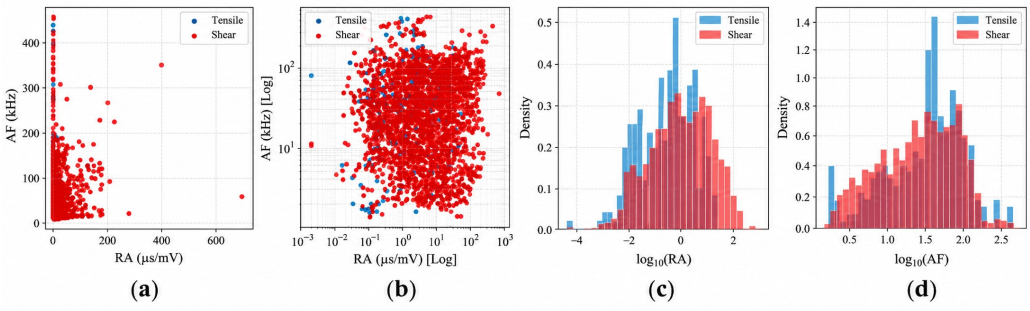


Figure 4. Distribution of the conventional RA–AF baseline in the event dataset used in this study: (**a**) scatter plot on linear coordinates; (**b**) scatter plot on logarithmic coordinates; (**c**) histogram of the RA distribution; and (**d**) histogram of the AF distribution.

3.4. Data Partitioning and Experimental Settings

The original dataset contained 2881 original AE events, including 262 tensile-associated and 2619 shear-associated events, corresponding to a class ratio of approximately 1:10. To achieve class balance while preserving the hierarchically locked partitioning strategy, a

global balancing–stratified locked partitioning procedure was adopted. First, the minority tensile class was balanced using random oversampling combined with waveform-level data augmentation. Starting from 262 independent tensile events, the number of tensile-event records was increased to 2619, matching the number of shear events and resulting in a balanced dataset containing 5238 entries. For the replicated tensile records, data augmentation was applied at the waveform level before CWT conversion, rather than at the CWT-image level. Specifically, three augmentation operations were used: random time shift along the temporal axis, Gaussian noise injection, and amplitude scaling with a random factor in the range of 0.8–1.2. The augmented waveforms were then transformed into CWT time–frequency images using the same processing pipeline as the original signals. The resulting balanced dataset comprised 5238 CWT records, including augmentation-derived records that do not represent additional independent physical AE events.

The original event ID was used as the grouping anchor throughout this procedure. All derived records associated with the same original event, including its augmented variants, shared the same global group ID and constituted an indivisible allocation unit during data partitioning. After class balancing, the dataset was divided by group ID using stratified random partitioning with a training/validation/test ratio of 7:1.5:1.5 while maintaining a 1:1 class ratio within each subset. The resulting partition was applied unchanged to all seven window durations following the hierarchically locked partitioning strategy described in Section 2.3.2.

After balancing and partitioning, the subset sizes were as follows: training set: 3666 records, comprising 1833 tensile-associated and 1833 shear-associated records; validation set: 786 records, comprising 393 tensile-associated and 393 shear-associated records; test set: 786 records, comprising 393 tensile-associated and 393 shear-associated records. The complete balanced CWT dataset, comprising 5238 records (2619 per class), was used for the class-conditional window-preference analysis. In contrast, the classification-performance evaluations were performed on the held-out test subset (786 recordsThe complete balanced CWT dataset, comprising 5,238 records (2619 per class), was used for the exploratory class-conditional window-preference analysis. In contrast, the classification-performance evaluations were performed on the held-out test subset (786 records, 393 per class). Together, these analyses characterize the influence of extraction-window duration on classification performance and architecture-dependent window preferences.

A separate model was trained for every architecture–window-duration combination, yielding $8 \times 7 = 56$ independently trained models. All models were trained under identical conditions. The batch size was 32, and the Adam optimizer was used with a learning rate of 1×10^{-3} and a weight-decay coefficient of 1×10^{-4}, The loss function was cross-entropy loss, implemented using CrossEntropyLoss. Class indices were used as supervision targets, with tensile events labeled as 0 and shear events labeled as 1. The raw logits produced by each model were passed directly to the loss function. A ReduceLROnPlateau scheduler was used with a patience of five epochs and a reduction factor of 0.5. The gradient norm was clipped at 1.0. The maximum number of training epochs was 100, and early stopping was applied when validation performance failed to improve for 15 consecutive epochs. Mixed-precision training was employed. To assess the stability of ensemble performance, the ResNet18 experiments were repeated with three random seeds (42, 123, and 2026). For each seed, the same event-paired 70/15/15 split was reused across all seven time windows, while model initialization and dropout were controlled by the corresponding seed. Performance is reported as the mean ± standard deviation across the three seeds. All experiments were conducted on a workstation equipped with an Intel Core i7-14700 processor and an NVIDIA GeForce RTX 5060 Ti GPU. The deep-learning models were implemented and trained using Python 3.9 and PyTorch 2.8.0, with CUDA

12.9 for GPU acceleration. The continuous wavelet transform was implemented using PyWavelets 1.6.0. The same computational environment and training configurations were maintained across all architecture–window-duration combinations to ensure consistency in the experimental comparisons.

Because the two classes were balanced after oversampling, accuracy was used as the principal classification-performance metric in the main results, supplemented by the area under the receiver-operating-characteristic curve (ROC-AUC) and the $F1$ score:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$\text{ROC} - \text{AUC} = P(\hat{p}(X^+) > \hat{p}(X^-)) \tag{13}$$

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{14}$$

where TP, TN, FP, and FN denote the numbers of true-positive, true-negative, false-positive, and false-negative predictions, respectively. In Equation (13), X^+ and X^- denote randomly selected positive and negative samples, respectively, and $\hat{p}(\cdot)$ denotes the positive-class probability predicted by the model.

For each test event i, let $\hat{y}_i^{\text{single}}$ denote the prediction produced by the single-window model selected according to validation-set accuracy $\hat{y}_i^{\text{ens}}$ denote the ensemble prediction, y_i denote the ground-truth label, and N denote the number of events in the test set. The event-level rescue rate, degradation rate, and net gain are defined as

$$\text{Rescue} = \frac{1}{N}\sum_{i=1}^{N} 1\left\{\hat{y}_i^{\text{single}} \neq y_i \wedge \hat{y}_i^{\text{ens}} = y_i\right\} \tag{15}$$

$$\text{Broken} = \frac{1}{N}\sum_{i=1}^{N} 1\left\{\hat{y}_i^{\text{single}} = y_i \wedge \hat{y}_i^{\text{ens}} \neq y_i\right\} \tag{16}$$

$$\text{NetGain} = \text{Rescue} - \text{Broken} \tag{17}$$

For reporting purposes, the rescue and degradation rates are expressed as event counts by multiplying them by N. This representation enables a direct event-by-event comparison between the selected single model and the ensemble.

To quantify the differential preference of each window duration for tensile and shear events, it is first necessary to determine the window under which each individual event is identified most reliably. For event i and architecture a, the subset of window durations under which the event is correctly classified is defined as

$$\mathcal{L}_i^{cor}(a) = L\left(L \in \mathcal{L} : \hat{y}_{a,L}\left(X_i^{(L)}\right) = y_i\right) \tag{18}$$

When $\mathcal{L}_i^{\text{cor}}(a) = \varnothing$ the event is excluded from the window-preference analysis. Within the subset of correctly classified windows, the event-level optimal window duration is defined as the window producing the highest posterior probability for the true class:

$$L_i^{\star}(a) = \arg\max_{L \in \mathcal{L}_i^{cor}(a)} p_{a,L}\left(y_i \mid X_i^{(L)}\right) \tag{19}$$

The number of class-c events assigned to window duration L as their event-level optimum is then defined as $n_c(L) = |i : y_i = c,\ L_i^{\star}(a) = L|$,The odds ratio (OR) is subsequently used to quantify the relative preference of window duration L for class c:

$$\text{OR}_c(L) = \frac{\frac{n_c(L)}{[N_c - n_c(L)]}}{\frac{n_{\bar{c}}(L)}{[N_{\bar{c}} - n_{\bar{c}}(L)]}} \tag{20}$$

where c is the target class, $\bar{c}$ is the reference class, and N_c and $N_{\bar{c}}$ are the total numbers of events in the two classes, respectively. A value of $\mathrm{OR_c(L)} > 1$ indicates that window duration L exhibits a statistical preference for class c. For a given architecture a, the class-conditional optimal window duration can then be extracted as $L^{\star}(c) = \arg\max_{L} \mathrm{OR}_c(L)$.

This quantity is used in Section 4.2 to characterize the architecture specificity and cross-architecture similarity of the optimal window durations.

The class-conditional window-preference analysis in Section 4.2 was conducted on the complete balanced dataset containing 5238 CWT records (2619 tensile-associated and 2619 shear-associated records), rather than exclusively on the independent test subset. These records include oversampled and augmented instances. Accordingly, the resulting OR distributions are used for exploratory characterization of model-dependent relative window preferences and should not be interpreted as independent-event estimates of generalization performance. Furthermore, in this binary formulation, the odds ratios obtained by exchanging the target and reference classes are reciprocal. Consequently, differences between the locations of their extrema do not, by themselves, demonstrate an architecture-independent physical mechanism.

3.5. Supplementary Signal-Level Detected-Event Duration Analysis

To examine whether model-derived window preferences correspond directly to the temporal extent of the detected AE events, an additional signal-level duration analysis was performed using the original continuous AE waveforms. This analysis was conducted independently of the classification models and CWT representations, consistent with the layer-wise design of MTS-AE.

The continuous waveforms were reprocessed using the same multi-scale RMS-envelope and adaptive-threshold procedure described in Section 2.2. The 0.2-ms RMS envelope was used as the principal detection envelope, and the local background statistics were evaluated within a 20-ms moving window. The major, medium, and minor adaptive thresholds were defined using coefficients of 3.5, 2.5, and 1.5, respectively, together with the corresponding global safeguard thresholds. Candidate detections satisfying the original event-merging criterion were treated as the same detected event.

For each merged event, the signal-level duration was defined as the temporal span from the earliest start to the latest end of the minor-threshold regions belonging to the same event group. The minor threshold was adopted for boundary determination because it preserves the lower-amplitude portions of the detected response while retaining consistency with the original three-level detection framework. If a merged group contained no minor-threshold region, the widest available detector-defined region was used. The fixed 2-ms waveform extraction window employed during the original dataset construction was retained only for event-validity checking and did not constrain the duration measured in this supplementary analysis.

The measured durations were divided into six intervals: (0,1], (1,2], (2,3], (3,4], (4,5], and (5,10] ms. Because the numbers of detected tensile- and shear-associated events were unequal, occurrence frequencies were normalized within each class. In the following discussion, "frequency distribution" therefore refers to the occurrence-frequency distribution of measured detected-event durations rather than to spectral frequency.

Reprocessing the complete original continuous-waveform collection yielded 265 tensile-associated detections and 2619 shear-associated detections. By contrast, the final classification dataset retained 262 tensile-associated and 2619 shear-associated original AE events. The supplementary duration analysis was conducted independently of the classification models and was intended to characterize the overall distribution of threshold-defined detected-event durations, rather than to provide a one-to-one reconstruction of the

classification records. Accordingly, the two analyses are reported using their respective statistical populations.

4. Results and Analysis

4.1. Overall Performance of Individual Models

Table 3 presents the classification accuracies of 56 architecture–window-duration configurations, comprising eight deep-learning architectures and seven extraction-window durations, evaluated on the held-out test subset ($N = 786$). The results provide an overview of the performance variations associated with window-duration selection across different network architectures.

Table 3. Test-set classification accuracy (%) of the eight architectures under the seven extraction-window durations, evaluated on the held-out test subset ($N = 786$). For each architecture, the window configuration with the highest test-set accuracy is shown in bold.

Architecture	0.5 ms	1 ms	2 ms	3 ms	4 ms	5 ms	10 ms
ResNet18	97.34	98.35	98.73	**98.98**	98.73	**98.98**	98.86
DenseNet121	96.83	98.60	**98.86**	98.73	98.73	98.60	98.73
EfficientNet-B0	94.92	97.72	97.97	**98.73**	98.60	98.60	97.46
MobileNetV2	97.84	97.59	97.84	**98.73**	98.48	98.35	98.10
KAN_Hybrid_Net	97.21	97.59	98.10	98.60	98.73	**99.11**	98.10
UKAN	95.56	96.70	96.45	96.45	96.57	97.21	**97.72**
Bi-Mamba2_row	86.55	89.72	88.45	93.15	**95.81**	88.20	90.36
Mamba2_row	83.76	80.33	87.06	88.96	**94.67**	92.77	91.75

The CNN architectures and KAN_Hybrid_Net generally fell within the high-performance range, with peak accuracies of at least 98.7%. UKAN achieved an intermediate performance level of approximately 97%. By contrast, the Mamba-based architectures exhibited an overall "increase followed by decrease" pattern as window duration increased and produced lower accuracies, ranging approximately from 84% to 96%. Among all 56 configurations, KAN_Hybrid_Net achieved the highest accuracy of 99.11% at 5 ms, highlighting the influence of window-duration selection on classification performance. The variation in performance with window duration for each architecture is shown in Figure 5.

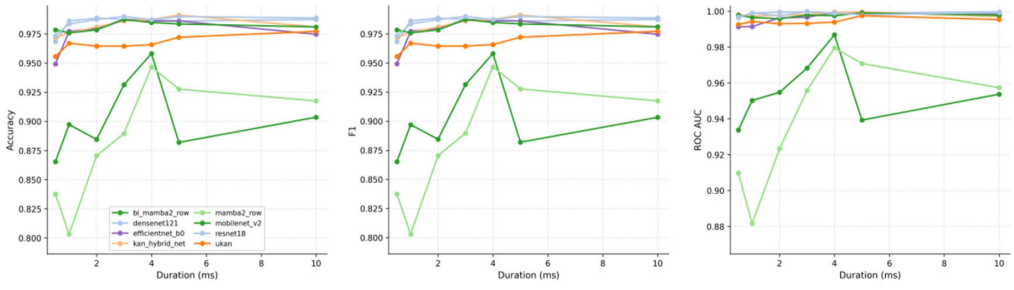


Figure 5. Performance curves of the eight candidate architectures across seven extraction-window durations in terms of accuracy, F1 score, and ROC-AUC, evaluated on the held-out test subset ($N = 786$). The CNN architectures and KAN_Hybrid_Net generally exhibit higher classification performance, whereas the Mamba-based architectures show lower overall accuracy and greater sensitivity to window duration.

4.2. Architecture-Specific Class-Conditional Window Preferences

To investigate the influence of network architecture on class-conditional window preferences, the odds-ratio (OR) metric defined in Equation (20) was applied to the eight

evaluated architectures. In contrast to the test-set evaluation reported in Section 4.1, this analysis used the complete balanced CWT dataset containing 5238 records (2619 per class), including oversampled and augmented instances. Table 4 summarizes the window durations exhibiting the strongest relative preferences for tensile- and shear-associated records within each architecture.

Table 4. Class-conditional optimal window durations determined using the OR metric on the complete balanced CWT dataset ($N = 5238$). T and S denote tensile and shear events, respectively. The optimal-window difference is the absolute difference between the preferred tensile and shear window durations within the same architecture.

Architecture	Optimal Window for T	Optimal Window for S	Optimal-Window Difference
ResNet18	3	10	7 ms
DenseNet121	3	2	1 ms
EfficientNet-B0	5	1	4 ms
MobileNetV2	3	5	2 ms
KAN_Hybrid_Net	2	10	8 ms
UKAN	2	5	3 ms
Mamba2_row	1	4	3 ms
Bi-Mamba2_row	10	4	6 ms

Table 4 reveals substantial architecture-dependent variation in the locations of class-conditional OR extrema. ResNet18 and KAN_Hybrid_Net exhibit relatively shorter preferred windows for tensile-associated records and longer preferred windows for shear-associated records, whereas the remaining architectures show different preference patterns. Therefore, the specific preferred durations cannot be generalized across the evaluated architectures.

Although the preferred windows of the two categories do not coincide in any of the eight architectures, the class-specific odds ratios are reciprocal in the present binary formulation. Consequently, this non-overlap should not be interpreted as independent evidence of a common physical fracture mechanism. Instead, the results characterize architecture-dependent relative preferences across the investigated window durations and motivate further investigation of complementary temporal representations through cross-window ensemble learning.

ResNet18 was selected for further descriptive analysis because it achieved the highest mean accuracy across the seven window durations among the evaluated architectures and the highest accuracy after cross-window ensemble fusion. Its class-conditional preferred-window distribution and corresponding OR statistics are presented in Table 5 and Figure 6a,b.

Table 5. Distribution of model-derived preferred window durations and class-conditional odds ratios for ResNet18 on the complete balanced CWT dataset ($N = 5238$; 2619 records per class).

L (ms)	$n_{tensile}$	n_{shear}	$OR_{tensile}(L)$	Class Preference
0.5	140	111	1.27	tensile × 1.27
1	151	147	1.03	tensile × 1.03
2	449	301	1.59	tensile × 1.59
3	722	235	3.86	tensile × 3.86
4	572	854	0.58	shear × 1.73
5	301	317	0.94	shear × 1.06
10	284	654	0.37	shear × 2.73

As shown in Table 5, ResNet18 exhibits the strongest relative preference for tensile-associated records at 3 ms (OR = 3.86) and for shear-associated records at 10 ms (OR = 2.73). The preference direction changes between 3 and 4 ms, indicating distinct class-conditional patterns across the investigated window durations. These findings suggest that the temporal information contributing to classification differs between the two categories under the ResNet18 architecture. The relationship between model-derived window preferences and measured AE event durations is further examined in Section 5.1.

Because event detection and window extraction are separate processes, a preferred extraction-window duration may extend beyond the threshold-defined duration of the detected AE response.

The signal-level detected-event duration distributions obtained using the procedure described in Section 3.5 are shown in Figure 6c. Among the tensile-associated detections, 206 of 265 events (77.74%) had durations within 1 ms, compared with 2014 of 2619 shear-associated detections (76.90%). The two categories exhibited substantial overlap, although shear-associated detections showed slightly higher proportions in the 1–3 ms range. These findings indicate that the model-derived 3-ms and 10-ms window preferences do not directly correspond to measured event durations, but reflect differences in the temporal context used for classification.

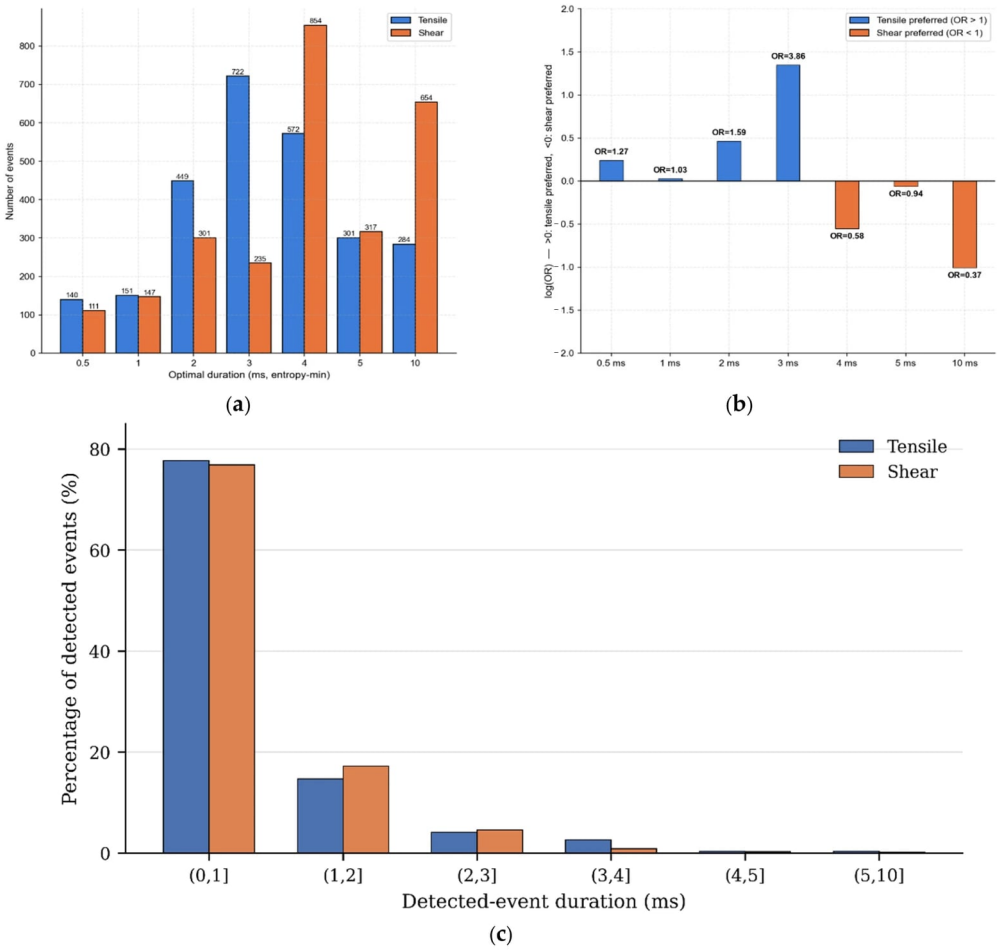


Figure 6. Model-derived temporal-window preferences and signal-level detected-event duration distributions. (**a**) Numbers of tensile- and shear-associated CWT records assigned to each record-level optimal window duration for ResNet18 on the complete balanced dataset; (**b**) class-preference strength at each candidate duration, expressed as the logarithm of the odds ratio; (**c**) class-normalized occurrence-frequency distributions of threshold-defined detected-event durations obtained by reprocessing the original continuous AE waveforms using the RMS-envelope adaptive-threshold detector.

4.3. Same-Architecture, Cross-Window-Duration Decision-Level Ensemble

Section 4.2 shows that the relative class-conditional window preferences vary across the investigated architectures and window durations. The supplementary signal-level analysis further indicates that these model-derived preferences cannot be explained solely by differences in threshold-defined detected-event durations. Instead, different extraction windows provide alternative temporal contexts for CWT representation and classification, potentially offering complementary discriminative information. Accordingly, this section evaluates a same-architecture, cross-window-duration decision-level ensemble. For each architecture, the softmax probabilities produced by seven independently trained window-specific models are combined using equal-weight averaging, as defined in Equation (8). The single-window reference model is selected from the seven candidates based on the highest validation-set accuracy and is subsequently evaluated on the test set to provide the baseline for comparison with the corresponding ensemble.

4.3.1. Overall Improvement

Table 6 summarizes the ensemble results for the eight architectures on the balanced test set (N = 786) in the original experimental run. The cross-window-duration ensemble achieved a positive net gain for all eight architectures. The improvement was generally larger for architectures with weaker single-window baselines. Mamba2_row, which had the lowest baseline accuracy, achieved the largest improvement of 2.41 percentage points, whereas DenseNet121 achieved a smaller improvement of 0.51 percentage points These two values (+2.41 and +0.51 percentage points) should be read against the seed-to-seed variation in the ensemble accuracy reported below (±0.26 percentage points); the small gains obtained by the strongest single-window baselines should therefore not be over-interpreted. Moreover, the number of rescued test records exceeded the number of degraded records for every architecture. In particular, DenseNet121 corrected four previously incorrect predictions without reversing any originally correct prediction in this test run. These results suggest that different extraction-window durations provide partially complementary classification information, allowing ensemble fusion to correct some errors made by the selected single-window models.

To assess the stability of this improvement, the ResNet18 experiments were repeated with three random seeds. The selected single-window model achieved a mean accuracy of 98.9848 ± 0.2198%, whereas the seven-window ensemble achieved 99.6616 ± 0.2642%. The corresponding rescued, degraded, and net-gain counts were 8/1/+7, 7/3/+4, and 5/0/+5 for seeds 42, 123, and 2026, respectively. Here, "rescued" denotes a test record incorrectly classified by the selected single-window model but correctly classified by the ensemble, whereas "degraded" denotes the opposite prediction reversal. Net gain is the difference between these counts. The positive net gain observed for all three seeds provides additional evidence that cross-window-duration ensemble fusion improves ResNet18 classification performance under the investigated evaluation setting. The architecture-specific results from the original run are visualized in Figure 7.

Table 6. Improvement achieved by multi-window ensemble fusion within each architecture on the test set ($N = 786$).

Architecture	Selected Window	Selected-Model Acc. (%)	Ensemble Acc. (%)	(\Delta) Acc. (%)	Rescued	Degraded	Net Gain
ResNet18	10 ms	98.86	99.75	+0.89	8	1	+7
DenseNet121	3 ms	98.73	99.24	+0.51	4	0	+4

Table 6. *Cont.*

Architecture	Selected Window	Selected-Model Acc. (%)	Ensemble Acc. (%)	(\Delta) Acc. (%)	Rescued	Degraded	Net Gain
EfficientNet-B0	10 ms	97.46	99.24	+1.78	15	1	+14
MobileNetV2	3 ms	98.73	98.98	+0.25	5	3	+2
KAN_Hybrid_Net	10 ms	98.10	98.73	+0.63	10	5	+5
UKAN	10 ms	97.72	98.86	+1.14	14	5	+9
Bi-Mamba2_row	4 ms	95.81	97.21	+1.40	17	6	+11
Mamba2_row	4 ms	94.67	97.08	+2.41	29	10	+19

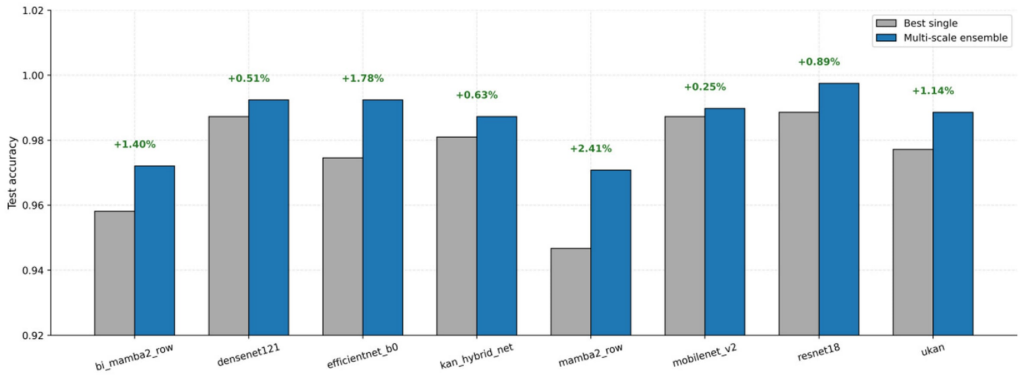


Figure 7. Multi-scale ensemble improvement for the eight architectures. Gray bars represent the test accuracy of the selected single model, and blue bars represent the accuracy of equal-weight fusion across the seven window durations. The absolute improvement is annotated above each pair of bars. All architectures exhibit a positive net gain.

To further characterize class-specific classification performance, F1-scores for both classes and balanced accuracy were calculated from the archived ResNet18 predictions. In the original experimental run, balanced accuracy increased from 98.86% for the selected 10-ms single-window model to 99.75% for the seven-window ensemble. Across the three-seed analysis, the corresponding mean balanced accuracies were 98.98 ± 0.22% and 99.66 ± 0.26%, respectively. The tensile-class F1-score increased from 98.87% to 99.75%, while the shear-class F1-score increased from 98.85% to 99.75%. These results indicate that the improvement achieved through ensemble fusion was reflected in both class-specific F1-scores under the investigated balanced-record evaluation setting.

4.3.2. Ablation Comparison of Ensemble Strategies and Ensemble Paths

Section 4.3.1 confirms the effectiveness of cross-window ensemble fusion. However, the ensemble design involves two orthogonal dimensions: how the predictions are aggregated, referred to as the ensemble strategy, and which models are aggregated, referred to as the ensemble path. This section separately examines these two dimensions to identify the source of the performance gain and determine the default configuration.

(1) Ensemble Strategy. Using ResNet18 as the representative architecture, the four strategies defined in Equations (8)–(11) were compared with the selected single-window baseline, as summarized in Table 7 for the original experimental run. To further assess the stability of the strategy comparison, three random seeds were evaluated. Equal-weight averaging, validation-accuracy weighting, hard voting, and confidence weighting achieved mean accuracies of 99.6616 ± 0.2642%, 99.6616 ± 0.2642%, 99.6616 ± 0.2642%, and 99.4501 ± 0.3194%, respectively. Thus, the first three strategies

achieved identical reported mean accuracies across the three seeds, whereas confidence weighting showed a slightly lower mean accuracy, although it still outperformed the selected single-window baseline These differences are small and lie within the seed-to-seed variation in the ensemble accuracy; they should therefore not be interpreted as evidence that one fusion rule is superior to another, the more so because this comparison is based on a single architecture (ResNet18) and a single event-paired data split. One possible limitation of confidence weighting is that a highly confident but incorrect prediction may receive excessive influence during fusion. For example, the challenging shear-associated sample shear_01081, discussed in Section 5.1, was misclassified with displayed softmax confidence scores of 1.00 under the 2-, 3-, and 4-ms windows. This example illustrates the potential risk of relying on prediction confidence, although it does not independently establish the cause of the observed performance difference among fusion strategies. Because equal-weight averaging requires no validation-dependent weighting coefficients or additional fusion hyperparameters and achieved competitive performance across the evaluated seeds, it was adopted as the default ensemble strategy.

Table 7. Test-set performance of ResNet18 using four ensemble strategies and the selected-single-model baseline.

Strategy	Accuracy (%)	ROC-AUC (%)
Selected single model	98.86	99.92
Equal-weight averaging	99.75	99.999
Validation-accuracy weighting	99.75	99.999
Hard voting	99.75	99.999
Confidence weighting	99.49	99.992

(2) Ensemble Path. To separate the relative contributions of window-duration diversity and architecture diversity, three ensemble paths were constructed:

1. Same-architecture, cross-window-duration ensemble: the seven window-specific models within each architecture were fused, producing eight ensemble instances.
2. Same-window-duration, cross-architecture ensemble: the eight architectures trained at each window duration were fused, producing seven ensemble instances.
3. Full ensemble: all 56 models were fused using equal-weight averaging, producing one ensemble instance.

The distributions of test-set accuracy for the three paths are presented in Figure 8.

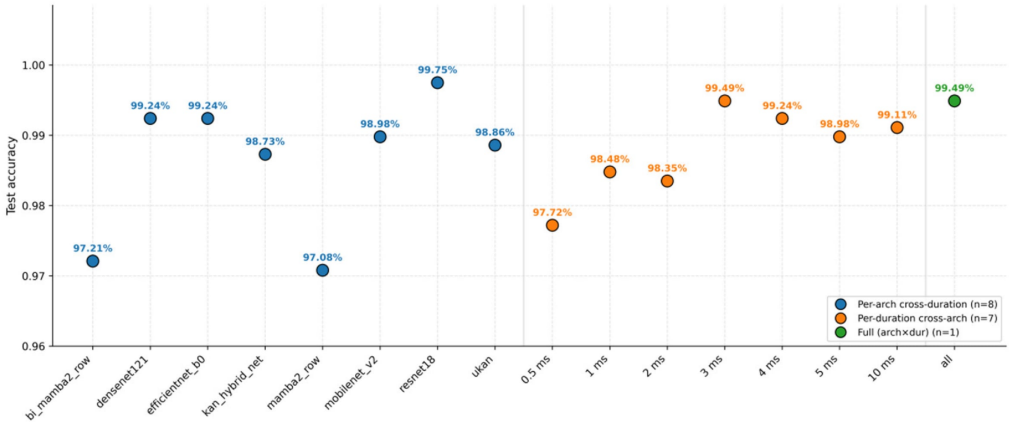


Figure 8. Comparison of test-set accuracy among three ensemble paths. Blue points represent same-architecture, cross-window-duration ensembles, with one point for each architecture (n = 8).

Orange points represent same-window-duration, cross-architecture ensembles, with one point for each window duration (n = 7). The green point represents the full equal-weight ensemble of all 56 models (n = 1).The vertical gray lines separate the three ensemble paths.

Figure 8 reveals three principal findings. First, the best-performing ensemble was obtained through the same-architecture, cross-window-duration path. The ResNet18 cross-window ensemble achieved a test accuracy of 99.75%, compared with 99.49% for both the best same-window, cross-architecture ensemble at 3 ms and the full 56-model ensemble. This margin is small (0.26 percentage points) and is of the same order as the seed-to-seed variation in the ensemble accuracy; it is therefore reported as an observation under the present evaluation setting rather than as a general claim that time-scale diversity is superior to architectural diversity. Second, increasing the number of ensemble members did not necessarily improve classification performance under the investigated evaluation setting. Although the full ensemble combined all 56 models, it did not outperform the seven-member ResNet18 cross-window ensemble in the original experimental run. This result suggests that the effectiveness of ensemble learning depends on the predictive contributions and complementarity of the participating models, rather than on model count alone. Third, the two ensemble paths exhibited different accuracy distributions. The performance variation among same-architecture, cross-window-duration ensembles was strongly associated with differences in the underlying architectures, particularly the relatively lower accuracies of the Mamba-based models. In contrast, the performance of same-window, cross-architecture ensembles varied with the temporal coverage used to construct their common CWT input representations. The relatively low performance at 0.5 ms and the stronger performance at intermediate window durations indicate that the available temporal context may influence the discriminative information captured across architectures. However, these results should be interpreted as representation-level effects rather than direct evidence of differences in intrinsic fracture-event duration.

The principal engineering implication is that, for a sufficiently strong architecture, representation-level diversity introduced through different extraction-window durations can provide ensemble benefits comparable to, or greater than, those obtained through architectural diversity under the investigated conditions. In the original experimental run, the seven-member ResNet18 cross-window ensemble achieved a marginally higher test accuracy than the full 56-model ensemble, while requiring substantially fewer model members. However, the reduction in model count should not be interpreted as a proportional reduction in inference latency, memory consumption, or system maintenance cost, since these quantities also depend on architectural complexity and implementation conditions. The full ensemble achieved 99.49% accuracy, compared with 97.21% and 97.08% for the same-architecture cross-window ensembles of Bi-Mamba2_row and Mamba2_row, respectively, indicating that combining different architecture families may remain beneficial when an individual architecture performs relatively poorly. Overall, selecting a high-performing architecture and applying cross-window ensemble fusion represents a promising deployment strategy, although its practical computational efficiency and real-time performance require direct benchmarking and validation under the target operating conditions.

4.3.3. Differences in Architectural Robustness

Evaluating an architecture solely on the basis of its peak accuracy can obscure an important engineering trade-off. An architecture that performs extremely well at one window duration but deteriorates sharply at other durations requires precise window tuning during deployment, thereby increasing system-design costs. To quantify this trade-

off, the accuracy spread across window durations and the mean window-duration accuracy are defined as

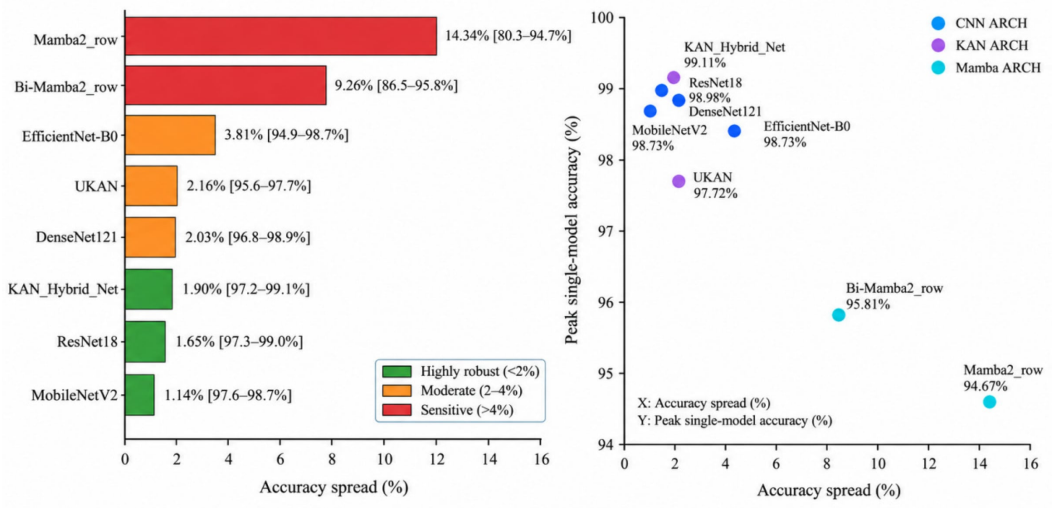
$$\text{Spread}(a) = \max_{L} \text{Acc}(a, L) - \min_{L} \text{Acc}(a, L), \overline{\text{Acc}}(a) = \frac{1}{|\mathcal{L}|} \sum_{L} \text{Acc}(a, L) \quad (21)$$

The accuracy spread reflects the sensitivity of each architecture to window-duration selection, whereas the mean accuracy summarizes its overall performance across the seven investigated durations.The corresponding results are summarized in Table 8.

Table 8. Joint robustness–accuracy assessment of the eight architectures on the held-out test subset ($N = 786$), derived from the single-model accuracies reported in Table 3. Architectures are ordered by increasing accuracy spread.

Architecture	Spread (%)	$\bar{\text{Acc}}$ (%)	σ_{Acc} (%)	Robustness Tier
MobileNetV2	1.14	98.13	0.38	High
ResNet18	1.65	98.57	0.54	High
KAN_Hybrid_Net	1.90	98.21	0.62	High
DenseNet121	2.03	98.44	0.66	Moderate
UKAN	2.16	96.66	0.62	Moderate
EfficientNet-B0	3.81	97.72	1.23	Moderate
Bi-Mamba2_row	9.26	90.32	2.95	Sensitive
Mamba2_row	14.34	88.47	4.76	Sensitive

The joint robustness–accuracy assessment presented in Figure 9, computed from the test-set accuracies in Table 3, identifies three distinct robustness tiers. MobileNetV2, ResNet18, and KAN_Hybrid_Net are classified as highly robust, with accuracy spreads below 2%. DenseNet121, UKAN, and EfficientNet-B0 fall within the moderate-robustness tier, with spreads ranging from 2% to 4%. In contrast, Bi-Mamba2_row and Mamba2_row exhibit substantially greater sensitivity to window duration, with spreads exceeding 4%. These results indicate that the CNN-based architectures generally maintain more stable classification performance across the investigated window durations, whereas the Mamba-based architectures exhibit greater performance variability. The pronounced window-duration sensitivity of the latter is discussed in Section 5.4.



(**a**) (**b**)

Figure 9. Comparison of window-duration robustness and peak classification accuracy among the eight architectures based on the held-out test subset ($N = 786$). (**a**) Robustness classification based

on accuracy variation across window durations. The horizontal axis represents the accuracy spread, and the architectures are categorized as highly robust, below 2%; moderately robust, 2–4%; or sensitive, above 4%. The accuracy range across window durations is shown in parentheses. (**b**) Robustness–accuracy plane. The horizontal axis represents the accuracy spread across window durations, and the vertical axis represents the best single-model accuracy. Marker categories denote the three architecture families, and marker colors distinguish the CNN, KAN, and Mamba architecture families.

4.4. Comparison with Established Architectures and Published Multi-Scale Methods

To further evaluate the proposed MTS-AE framework in the context of existing methods, comparisons were conducted at two levels: a controlled comparison among established network architectures using the present dataset, and a literature-based comparison with published studies on AE classification and multi-scale signal analysis. At the architecture level, Tables 3 and 6 provide complementary evaluations of the eight representative architectures from the CNN, KAN, and Mamba families. Table 3 compares their single-model accuracy across the seven extraction-window durations on the held-out test subset, while Table 6 reports the accuracy improvements achieved through cross-window ensemble fusion on the same subset. In particular, ResNet18 achieved an ensemble accuracy of 99.75%, compared with 98.86% for its validation-selected single-window model. Moreover, as shown in Figure 8, this seven-model cross-window ensemble exceeded the accuracy of the full 56-model ensemble (99.49%) under the present evaluation conditions. These results suggest that cross-window prediction diversity can improve classification performance across different architecture families under the adopted evaluation protocol, although the magnitude of improvement varies among architectures.

To further clarify the relationship between MTS-AE and previously published approaches, representative studies addressing fracture-mode identification, waveform sequence length, and multi-scale feature learning were examined. Their datasets, model inputs and outputs, time-scale treatments, and main findings are summarized in Table 9.

The studies summarized in the comparison table demonstrate that temporal information has been investigated through different methodological approaches. Ren et al. examined parameter-based and waveform-based representations for tensile–shear AE classification, while Song et al. investigated the influence of waveform sequence length on fracture-stage identification. Bai et al. employed multi-scale clipping fusion for vibration-signal augmentation and fault diagnosis, whereas Guo et al. used multi-scale convolutional feature extraction for composite-material AE damage classification. Unlike these approaches, MTS-AE systematically varies the extraction-window duration of the same AE event while maintaining consistent event pairing, data partitioning, and CWT preprocessing parameters. The resulting cross-window predictions are subsequently combined through decision-level ensemble fusion. The experiments conducted with eight architectures support the effectiveness of this strategy under the present evaluation protocol. However, because the published studies involve different datasets, tasks, and experimental settings, their reported performance metrics are not directly comparable with those of MTS-AE. Therefore, this literature-based comparison establishes methodological distinctions rather than quantitative superiority over the published approaches.

Table 9. Comparison of published AE classification and multi-scale signal analysis methods with the present study.

Study	Dataset and Task	Model Input → Output	Time-Scale Treatment	Main Findings
Ren et al. [23]	Red sandstone; tensile/shear AE identification	AE parameters or waveform-based images → two AE classes	Different signal representations; no systematic event-window scan	Decision-tree accuracy up to 88%; CNN-based models performed better
Song et al. [26]	Sandstone under true triaxial loading; fracture-stage identification	AE waveform sequences → three fracture stages	Six sequence lengths: 1, 2, 3, 5, 10, and 15 waveforms	Maximum accuracy of 90.4% at a sequence length of 10
Bai et al. [35]	Rolling-bearing vibration; fault diagnosis	Multi-scale clipped signals and STFT images → fault categories	Multi-scale clipping fusion for data augmentation	Improved fault classification using multi-scale information
Guo et al. [36]	Composite-material AE; damage classification	Raw AE sequences or frequency-domain sequences → three damage classes	InceptionTime with multi-scale convolutional feature extraction	Classification accuracy of approximately 99%
Present study	Marble AE under Brazilian splitting and uniaxial compression; tensile-/shear-dominated failure classification	Seven paired CWT representations → two AE categories	Seven event-extraction windows from 0.5 to 10 ms; decision-level fusion	ResNet18 ensemble accuracy of 99.75%; class-conditional window preferences examined

Note: Results from previously published studies were obtained using different datasets, classification tasks, and evaluation protocols. Their reported accuracies are provided for descriptive comparison and should not be interpreted as direct same-dataset performance benchmarks. In particular, waveform sequence length, signal clipping scale, neural-network receptive-field scale, and individual-event extraction-window duration represent different aspects of temporal information.

5. Discussion

5.1. Interpretation of Window-Duration Preferences from Signal- and Representation-Level Perspectives

The ResNet18 results in Section 4.2 identify the strongest relative window preferences at 3 ms for tensile-associated records and 10 ms for shear-associated records. In contrast, the supplementary signal-level analysis shows that approximately 77% of the detected events in both categories have durations within 1 ms, with substantial overlap between their duration distributions. This difference suggests that the preferred extraction windows reflect the broader temporal information available to the classifier rather than simply the duration of the detected AE response. The relationship between extraction-window duration and event characteristics therefore needs to be examined from both signal-level and representation-level perspectives.

This distinction can be understood from the layered design of MTS-AE. At the signal layer, adaptive-threshold detection defines the principal recorded AE response and its operational temporal boundaries. At the representation layer, the extraction window determines how much pre-peak, post-peak, decay, propagation-related, and background information is mapped into the CWT representation. Consequently, the temporal context available for model discrimination can extend beyond the threshold-defined duration of the

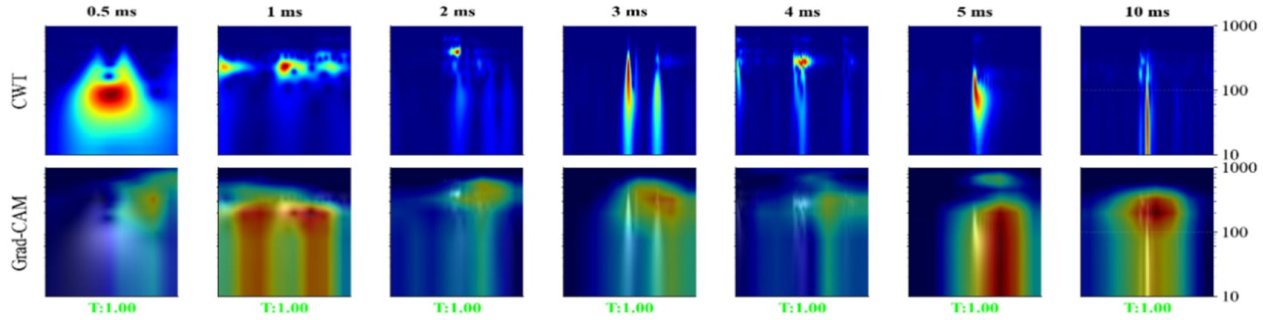
detected event itself. This provides a representation-level explanation for why classification performance can remain sensitive to extraction-window duration even when the measured duration distributions of the two categories overlap substantially.

Recorded AE signals are additionally influenced by wave propagation, attenuation, sensor response, and possible mode conversion. These factors further reinforce the need to distinguish the duration of a detector-defined recorded response from the broader temporal context represented in the extracted waveform. Accordingly, the 10-ms preference observed for shear-associated records under ResNet18 suggests that broader temporal context can contribute useful discriminative information under the present experimental configuration.

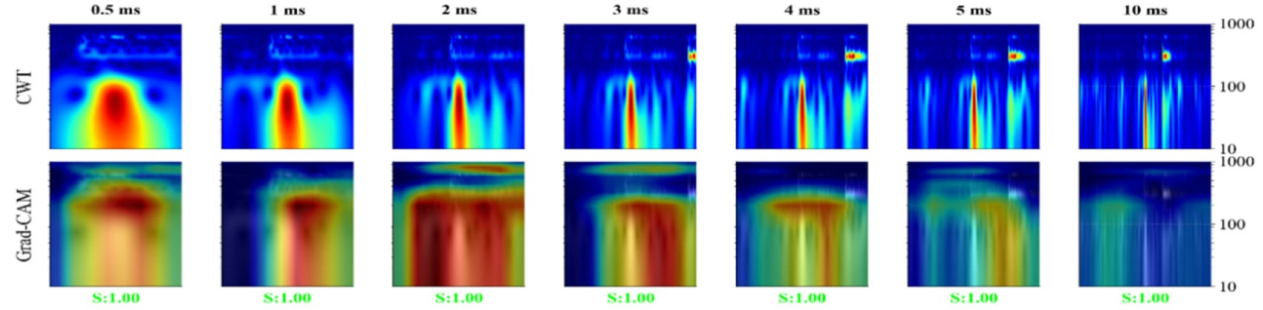
The distinction between signal-level duration and representation-level temporal context can also be examined through Grad-CAM. The Grad-CAM method proposed by Selvaraju et al. [37] was used to visualize the window-dependent discriminative regions emphasized by ResNet18. The tensile-associated sample tensile_00747 and shear-associated sample shear_00034 were deliberately selected because both were correctly classified under all seven window durations with softmax confidence scores of at least 0.95. These consistently correct, high-confidence samples provide clear examples for visualizing changes in model attention, although they do not necessarily represent the full variability of AE signals within each category. As shown in Figure 10a, the attention for tensile_00747 remains relatively concentrated around the principal time–frequency structures, whereas Figure 10b shows that the attention distribution for shear_00034 changes across window durations and is redistributed across different portions of the time–frequency representation. These observations demonstrate how changes in temporal coverage influence the regions emphasized by the classifier, providing a visual explanation for its window-dependent behavior.

To complement these consistently correct examples, the more challenging shear-associated sample shear_01081 was examined in Figure 10c. This sample was correctly classified under the 0.5-, 1-, and 5-ms windows but was misclassified as tensile under the 2-, 3-, 4-, and 10-ms windows. The incorrect predictions at 2–4 ms were associated with displayed tensile confidence scores of 1.00, while the 10-ms model also misclassified the sample with a tensile confidence score of 0.67. This non-monotonic behavior demonstrates that extending the extraction window does not simply add more of the same information; instead, changing temporal coverage alters the organization and relative salience of the time–frequency features available to the network. Notably, although ResNet18 exhibits its strongest aggregate relative shear preference at 10 ms, this individual shear-associated sample is still misclassified at that duration. The result therefore provides a concrete sample-level illustration of representation dependence and offers qualitative motivation for the cross-window ensemble strategy. It also illustrates that high softmax confidence does not necessarily imply prediction correctness.

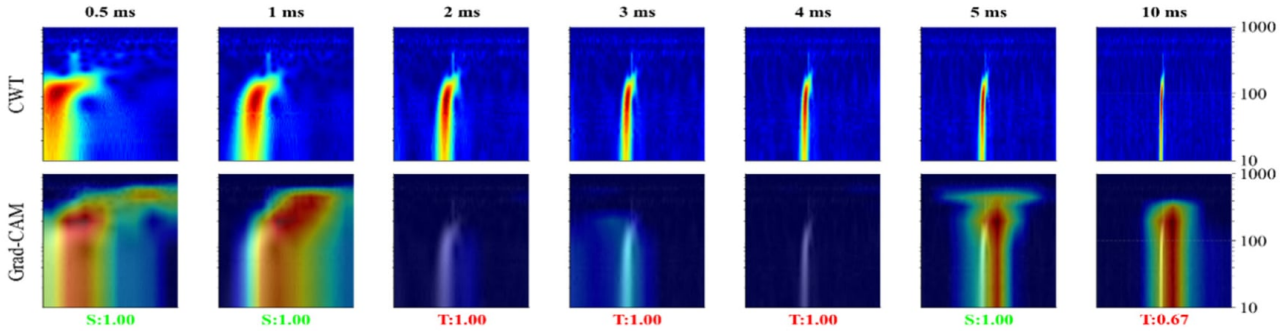
Overall, the signal-level duration analysis and Grad-CAM visualizations support a clear distinction between detector-defined event duration and representation-level temporal context. Extraction-window duration is therefore neither merely a preprocessing setting nor a direct surrogate for intrinsic fracture-event duration. Rather, it should be regarded as a model-dependent representation-design parameter that determines the temporal context mapped into the time–frequency domain. The preferred window within a given candidate set may consequently depend jointly on signal content, propagation and acquisition characteristics, time–frequency representation, and model architecture. The limitations associated with the available signal representations and event independence are further discussed in Section 5.4.



(**a**) Clean tensile event tensile_00747, correctly classified as tensile under all window durations



(**b**) clean shear event shear_00034, correctly classified as shear under all window durations



(**c**) cross-window reversal event shear_01081, exhibiting window-dependent misclassification

Figure 10. CWT time–frequency representations (**upper row**) and corresponding Grad-CAM attention maps (**lower row**) for three selected AE samples analyzed using ResNet18 under seven extraction-window durations (0.5, 1, 2, 3, 4, 5, and 10 ms). (**a**) Tensile-associated sample tensile_00747 and (**b**) shear-associated sample shear_00034, both deliberately selected because they were correctly classified under all seven windows with softmax confidence scores of at least 0.95; (**c**) challenging shear-associated sample shear_01081, correctly classified at 0.5, 1, and 5 ms but misclassified at 2, 3, 4, and 10 ms. Green labels indicate correct predictions, red labels indicate incorrect predictions, and the displayed numerical values represent the predicted-class softmax confidence scores.The color scale represents the normalized CWT magnitude (energy), with dark blue indicating lower values and red indicating higher values.

5.2. Cross-Window Error Structure of ResNet18

To characterize the consistency of misclassifications produced by the same architecture at different time scales, ResNet18 is used as a representative model. Let the set of events misclassified under window duration L be defined as

$$E_L = i : \hat{y}_L\left(X_i^{(L)}\right) \neq y_i, L \in \{0.5, 1, 2, 3, 4, 5, 10\}\text{ms} \quad (22)$$

The overlap between the error sets obtained under any two window durations is quantified using the Jaccard similarity coefficient:

$$J(E_L, E_{L'}) = \frac{|E_L \cap E_{L'}|}{|E_L \cup E_{L'}|} \tag{23}$$

A symmetric similarity matrix is constructed accordingly, as shown in Figure 11. Excluding the diagonal elements, which are necessarily equal to 1.00, the 21 unique off-diagonal window pairs generally exhibit low Jaccard similarities. Their mean is approximately 0.11, their median is only 0.07, and most values lie between 0.00 and 0.21. These results indicate that the sets of events misclassified under different window durations are only weakly correlated.

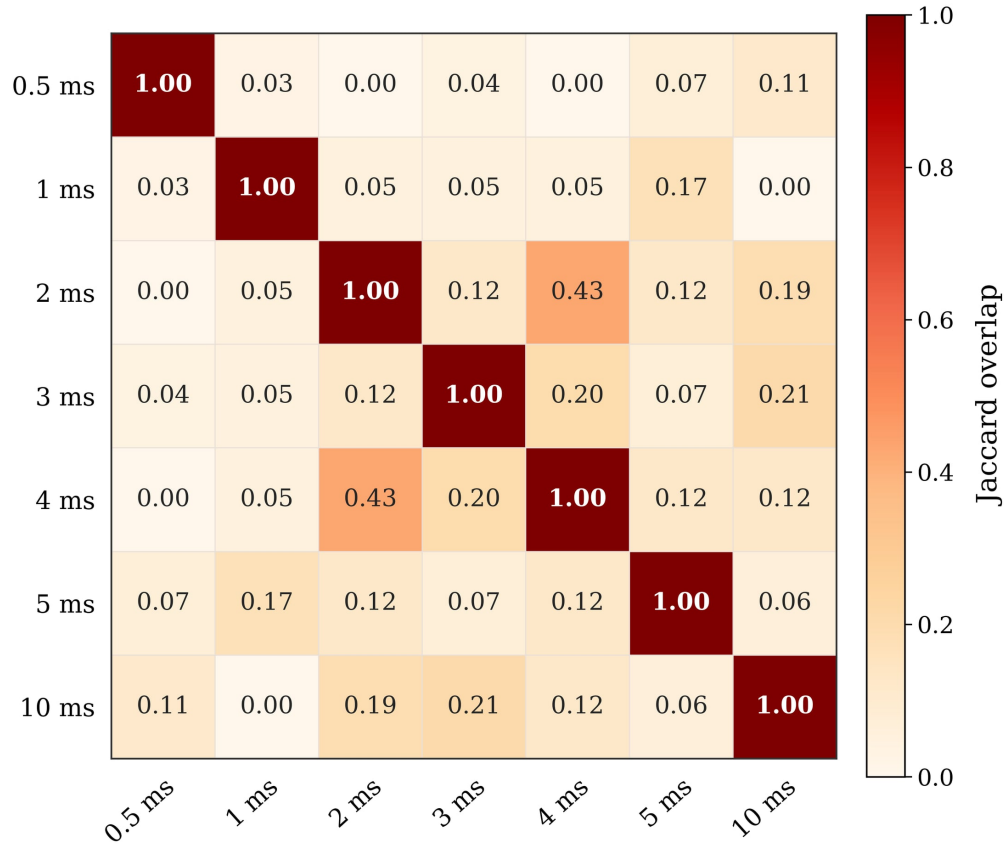


Figure 11. Jaccard similarity matrix of the event sets misclassified by ResNet18 under different window durations. The horizontal and vertical axes represent window durations in milliseconds. Each matrix element is the ratio between the intersection and union of the error sets obtained under the corresponding pair of windows. The diagonal elements are 1.00, and darker shading indicates greater error overlap.

A closer examination reveals several local patterns. The Jaccard similarities between the 0.5 ms window and the other six windows range from 0.00 to 0.11, with a row mean of 0.04, the lowest in the entire matrix. This finding reflects the fundamental difference in time–frequency resolution between the extremely short window and the longer windows. Difficult samples near the decision boundary under the 0.5 ms window are rarely shared with those under other durations. Some local consistency appears among intermediate windows. The Jaccard similarity between the 2 and 4 ms windows reaches 0.43, while those between 3 and 4 ms and between 3 and 10 ms are 0.20 and 0.21, respectively. Nevertheless, even the adjacent 2 and 3 ms windows have a similarity of only 0.12. Thus, proximity in window duration does not directly imply consistency in misclassification. Widely separated

windows share almost no errors. The Jaccard similarity is 0.00 between the 1 and 10 ms windows, 0.06 between the 5 and 10 ms windows, and only 0.11 between the 0.5 and 10 ms windows. These findings demonstrate that short and long windows capture substantially different difficult samples near the classification boundary.

The observed error structure highlights the complementarity of different extraction-window durations. First, the misclassified records vary across windows, indicating that no single duration consistently resolves all difficult cases. Second, even within the same network architecture, changing the temporal coverage produces models with substantially different error patterns. This diversity provides an important basis for cross-window ensemble learning, as predictions from different temporal representations can compensate for errors made by individual models. The Jaccard analysis thus offers further evidence supporting the effectiveness of window-duration diversity in the proposed ensemble framework.

5.3. Scenario-Specific Recommendations for Engineering Deployment

Considering the window-duration preferences identified in Section 4.2, the ensemble results presented in Section 4.3, and the robustness diagnosis reported in Section 4.3.3, the event-extraction window should not simply adopt the commonly used default values of 1 or 2 ms. Instead, it should be designed as an engineering parameter governing the alignment between signal characteristics and model architecture. Based on these findings, Table 10 summarizes several candidate configurations for different monitoring requirements, considering classification accuracy, window-duration robustness, and model complexity.

Table 10. Scenario-specific deployment recommendations. The reported single-model and ensemble accuracies are the held-out test-set values ($N = 786$) given in Table 6. The ensemble parameter count refers to the total number of parameters in the seven window-specific models.

Deployment Scenario	Recommended Configuration	Single-Model Accuracy	Ensemble Accuracy	Principal Rationale
High-accuracy research	ResNet18 with seven-window ensemble	98.86%	99.75%	Accuracy prioritized and sufficient computing resources available
Edge or embedded monitoring	UKAN with seven-window ensemble	97.72%	98.86%	Total ensemble size of only approximately 1.30 M parameters
Safety-critical monitoring	DenseNet121 with ensemble fusion	98.73%	99.24%	No degraded events: four rescued and zero degraded
Extremely high real-time requirement	MobileNetV2 single model at 3 ms	98.73%	98.98%	Highest robustness, with a spread of 1.14%

Three factors should be evaluated before deployment. First, window selection should account for both the target AE category and the model architecture, because the observed class-conditional preferences vary across the evaluated architectures. Second, the robustness of the architecture to window-duration perturbations, quantified by the accuracy spread, determines whether precise window tuning is required. Third, the available parallel-computing resources determine whether a multi-scale ensemble can be used. For monitoring scenarios dominated by tensile fracture, such as tensile-stress regions in brittle rock or early-age shrinkage cracking in concrete, a 2–3 ms single-window model may be used as the initial deployment configuration. A seven-window ensemble can then be added when computational resources permit. For monitoring scenarios potentially dominated by shear fracture,

such as fault zones and deep underground mining, a window of approximately 10 ms may be considered as an initial candidate based on the present experimental findings, subject to further validation under the target monitoring conditions. When the dominant fracture mode is unknown in advance, a same-architecture, cross-window-duration ensemble can be deployed directly to accommodate both tensile and shear events. However, the proposed deployment strategy has not yet been evaluated under independent field conditions, and its real-time feasibility and robustness in practical monitoring applications remain to be validated.

5.4. Transferability of the MTS-AE Method

The potential applicability of MTS-AE to other monitoring tasks is supported by its modular design, which separates event detection, multi-window extraction, and time–frequency representation into three coordinated processing levels. At the signal layer, the method assumes a three-component temporal structure consisting of a transient response, a decay phase, and background noise. This structure is broadly present in AE applications involving concrete, composite materials, and rotating-machinery fault monitoring. The multi-scale RMS-envelope and three-level adaptive-threshold mechanisms can therefore be adapted to these fields. At the event layer, the principle of reusing a global event ID across time scales is a standardized methodological operation. Its logic is independent of the specific fracture-mode labels, downstream architecture, and task objective. It can therefore serve as a general data-organization procedure for window-duration sensitivity analysis. At the representation layer, locking preprocessing parameters across windows—including the mother wavelet, frequency axis, and image resolution—is fundamentally an experimental-design strategy for controlling confounding variables. This principle is not strongly tied to a specific representation. When alternative time–frequency mappings such as the short-time Fourier transform, Mel spectrograms, or synchro squeezed wavelet transforms are used, cross-window comparability can be reconstructed by locking the corresponding parameters within the new representation system. The principal methodological value of MTS-AE is therefore not restricted to the particular seven-window dataset examined in this study. Rather, it provides a standardized analytical framework through which the effect of window duration can be rigorously separated from confounding factors such as perturbations in data partitioning and inconsistencies in preprocessing. Regarding model architecture, Mamba2_row and Bi-Mamba2_row belong to the selective state-space model family, which models sequential dependencies through input-dependent state transitions. Unlike CNN-based architectures that directly exploit local spatial relationships using two-dimensional convolutional kernels, the Mamba-based implementations in this study process CWT images through row-wise sequential scanning. Although this transformation preserves the ordering of features within individual rows, it does not explicitly maintain all local spatial relationships between adjacent rows. Consequently, the original two-dimensional relationships between temporal and frequency components may become less accessible to the sequential modeling mechanism, potentially limiting the extraction of localized time–frequency patterns that are important for AE fracture-mode identification.

This representational difference may partly explain the relatively lower classification accuracy and greater window-duration sensitivity of Mamba2_row and Bi-Mamba2_row observed in Tables 3 and 8. Moreover, because changing the extraction window duration modifies the temporal distribution of discriminative features in the CWT representation, the effectiveness of sequential feature modeling may vary across window durations. Nevertheless, the present experiments did not independently isolate the influence of the scanning strategy from other architectural characteristics. Therefore, the observed performance differ-

ences should be interpreted as potentially associated with the adopted representation and scanning mechanism rather than as an inherent limitation of Mamba-based architectures.

The principal findings of this study—the observed influence of window duration across the investigated architectures and the positive net gain produced by cross-window ensemble fusion—have internal validity under the controlled experimental conditions. Their broader generalization, however, remains constrained by three boundaries: dataset scale, labeling assumptions, and methodological configuration. Regarding dataset scale, the minority-class balancing procedure combined random oversampling with waveform-level data augmentation applied to the tensile events before CWT conversion. The adopted augmentation operations—time shift, Gaussian noise injection, and amplitude scaling—were intended to increase the variability of the minority-class samples and reduce the model's dependence on fixed waveform alignment, noise conditions, and absolute signal amplitude. The use of consistent record-level partitions across the seven window durations helped maintain comparability among the model evaluations. Since augmentation-derived CWT records do not represent additional independent physical AE events, the present results are interpreted within the balanced-record evaluation setting. Further assessment using a non-oversampled test set with verified original-event identities would provide additional evidence of event-level generalization.

Generalization to multiple lithologies, loading conditions, and acquisition configurations therefore requires validation using larger, multi-source datasets. Nevertheless, the experimental-type-based labeling strategy described in Section 3.1 has certain limitations. In particular, secondary tensile cracks may occur during the post-peak stage of uniaxial compression. Such complex fracture behavior, in which secondary tensile cracking coexists with a dominant shear mechanism, is forcibly assigned to a single category under the binary-label framework and may introduce label noise. Moreover, a binary tensile/shear classification does not cover finer-grained damage modes encountered in engineering materials, such as matrix cracking and fiber breakage. It may also obscure the discriminative characteristics of events involving coupled fracture mechanisms. Regarding the methodological configuration, the conclusions are specifically based on CWT time–frequency-image inputs and supervised deep learning. The time–frequency resolution and mother-wavelet parameters impose implicit constraints on the findings. Transfer to alternative approaches, including STFT representations, Mel spectrograms, one-dimensional convolutional networks, and self-supervised pretraining, must therefore be examined independently.

An additional limitation concerns the interpretation of the supplementary signal-level duration analysis. The reported event durations are operational quantities defined by the RMS-envelope adaptive-threshold detector rather than direct measurements of fracture-source time. Their numerical values may therefore depend on the envelope scale, adaptive-threshold parameters, event-merging criterion, wave propagation, attenuation, sensor response, and possible mode conversion. Moreover, reprocessing the complete raw-waveform collection yielded 265 tensile-associated detections, whereas 262 tensile events were retained in the final classification dataset. The supplementary duration analysis is therefore treated as an independent signal-level validation of the temporal-distribution tendency rather than as a one-to-one reconstruction of every classification record. Future work should explicitly preserve detector start/end boundaries together with the global event IDs used for multi-window extraction, enabling direct event-by-event linkage between measured duration, time–frequency representation, and model response.

6. Conclusions

This study addresses the long-standing reliance on empirical settings for event extraction-window duration in acoustic emission (AE) fracture-mode identification, and

establishes a research framework centered on multi-time-scale representation, class-conditional window-duration preference analysis, and cross-window-duration decision-level ensembles. The main conclusions are as follows.

(1) A multi-time-scale acoustic emission analysis method (MTS-AE) is proposed, providing a standardized procedure for controlled comparison of window-duration effects. At the signal layer, multi-scale RMS envelopes and a three-level adaptive threshold are used for event detection. At the event layer, the same physical event under seven window durations is strictly paired by a global event ID, and data-partition disturbance is controlled through hierarchical locked partitioning. At the representation layer, the CWT mother wavelet, frequency axis, and image resolution are fixed. This design separates window-duration effects from potential confounding factors such as event detection, preprocessing, and data partitioning, thereby making classification performance comparable across different time scales.

(2) The analysis of eight deep-learning architectures under seven candidate window durations reveals architecture-dependent class-conditional window preferences. Within the investigated candidate set, the durations corresponding to the extrema of the class-specific odds ratios differ between the tensile-associated and shear-associated categories across all eight architectures. However, because the class-specific odds ratios are reciprocal in the present binary formulation, this non-overlap should not be interpreted as independent evidence of a universal physical relationship. For ResNet18, the strongest relative preferences occur at 3 ms for tensile-associated records (OR = 3.86) and 10 ms for shear-associated records (OR = 2.73). These descriptive results were obtained from the complete balanced CWT dataset, which includes oversampled and augmented records. Supplementary signal-level analysis of the original continuous AE waveforms revealed substantial overlap in the threshold-defined detected-event duration distributions, with 206 of 265 tensile-associated detections (77.74%) and 2014 of 2619 shear-associated detections (76.90%) occurring within 1 ms. Therefore, the observed window preferences should be interpreted as model-dependent representation characteristics within the investigated experimental conditions rather than universal physical properties of individual fracture events.

(3) Based on the complementarity of models with different window durations on difficult samples, a same-architecture cross-window-duration decision-level ensemble scheme is constructed. This scheme achieves positive net gains on all eight architectures, and the test accuracy of ResNet18 increases from 98.86% to 99.75%. In the supplementary three-seed analysis (42, 123, and 2026), the selected single-window model achieved a mean test accuracy of 98.98 $\pm$ 0.22%, whereas the seven-window ensemble achieved 99.66 $\pm$ 0.26%. In the controlled setting of this study, the cross-window-duration ensemble of ResNet18 performs slightly better than both the best cross-architecture ensemble and the full ensemble of 56 models (both 99.49%); this margin of 0.26 percentage points should not be over-interpreted, and the result indicates that time-scale diversity can serve as an effective ensemble dimension for improving classification performance under the investigated conditions. For strong architectures with high window-duration robustness, the strategy of "strong architecture + multi-window parallel inference" can achieve high-accuracy identification with fewer model members.

(4) Several limitations should be kept in mind when interpreting these results. First, the balanced dataset was obtained by random oversampling of the tensile-associated class combined with waveform-level data augmentation, so the held-out test subset also contains augmentation-derived records. Because these records are derived from a limited number of independent tensile events, they do not represent additional independent physical AE events, and the reported record-level accuracies therefore

characterize performance on a balanced record set rather than event-level generalization to new AE data. Second, data partitioning was performed at the event-group level and was not stratified by specimen, so records belonging to the same specimen may occur in both the training and the test subsets; an independent specimen-level validation has therefore not yet been established. In addition, the class labels follow the experimental loading condition rather than an independent verification of the microscopic fracture mechanism of each event. Accordingly, the accuracy values reported here should be regarded as internally valid under the investigated experimental configuration, and their transfer to new AE datasets should be verified by specimen-held-out validation with test records that contain no augmented instances.

Author Contributions: Methodology, S.Y. and Q.J.; Software, Q.J.; Validation, C.Y. and L.W.; Formal analysis, S.Y.; Investigation, Z.C.; Resources, X.L.; Data curation, C.Y. and X.L.; Writing—original draft, S.Y.; Writing—review & editing, C.L.; Visualization, Z.C.; Project administration, C.L. and L.W.; Funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge the financial support of the State Key Laboratory of Disaster Prevention and Ecology Protection in Open-pit Coal Mines Open Projects Fund (Project NO. LTQZ-2026-202), Inner Mongolia Autonomous Region Science and Technology Program (Project NO. 2025YFHH0020), National Natural Science Foundation of China (Project NO. 52304265), Natural Science Foundation of Jiangsu Province (Project NO. BK20221121) and the Open Fund Project of State Key Laboratory of Safe Mining of Deep Coal and Environmental Protection (No. 2026YB012).

Data Availability Statement: The data that support the findings of this study are available from the corresponding authors upon reasonable request. Data access requests may be submitted via email to liuc_rm@cumt.edu.cn or leiwang@cumt.edu.cn.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ou, J.; Wang, E.; Wang, X.; Wang, B.; Zhu, G. Acoustic emission characteristics and damage evolution of different rocks under uniaxial compression conditions. *Sci. Rep.* **2024**, *14*, 4179. [CrossRef] [PubMed]
2. Khan, M.; He, X.; Song, D.; Li, Z.; Tian, X. Predicting deformation kinetics and fractures propagation in coal-rock masses using acoustic emission testing. *Int. J. Coal Sci. Technol.* **2025**, *12*, 28. [CrossRef]
3. Zhang, Q.B.; Zhao, J. A review of dynamic experimental techniques and mechanical behaviour of rock materials. *Rock Mech. Rock Eng.* **2014**, *47*, 1411–1478. [CrossRef]
4. Du, K.; Li, X.; Tao, M.; Wang, S. Experimental study on acoustic emission (AE) characteristics and crack classification during rock fracture in several basic lab tests. *Int. J. Rock Mech. Min. Sci.* **2020**, *133*, 104411. [CrossRef]
5. Wang, Y.; Han, J.; Li, C. Acoustic emission and CT investigation on fracture evolution of granite containing two flaws subjected to freeze–thaw and cyclic uniaxial increasing-amplitude loading conditions. *Constr. Build. Mater.* **2020**, *260*, 119769. [CrossRef]
6. He, M.; Miao, J.; Feng, J. Rock burst process of limestone and its acoustic emission characteristics under true-triaxial unloading conditions. *Int. J. Rock Mech. Min. Sci.* **2010**, *47*, 286–298. [CrossRef]
7. Wang, C.; Hou, X.; Liu, Y. Three-dimensional crack recognition by unsupervised machine learning. *Rock Mech. Rock Eng.* **2021**, *54*, 893–903. [CrossRef]
8. Wu, Y.; Yin, T.; Zhuang, D.; Chen, Y.; Ma, J.; Guo, W. Initial and unstable fracture toughness of quartz-diorite rock under mixed mode I/II loading: Experimental and numerical investigation. *Theor. Appl. Fract. Mech.* **2023**, *127*, 104079. [CrossRef]
9. Khadivi, B.; Heidarpour, A.; Zhang, Q.; Masoumi, H. Characterizing the cracking process of various rock types under Brazilian loading based on coupled Acoustic Emission and high-speed imaging techniques. *Int. J. Rock Mech. Min. Sci.* **2023**, *168*, 105417. [CrossRef]
10. Heng, Z.; Xu, T.; Konietzky, H.; Zhu, W.; Ranjith, P. New insights into the continuous-discontinuous failure characteristics of granite under Brazilian splitting test conditions using acousto-optic-mechanical (AOM) method. *Int. J. Rock Mech. Min. Sci.* **2025**, *187*, 106041. [CrossRef]
11. Zhou, T.; Chen, J.; Wang, F.; Zhou, C. Uncovering Failure and Cracking Behaviors of Double-Flawed Sandstone Under Compressive-Shear Loading Using Digital Image Correlation (DIC) Technique. *Rock Mech. Rock Eng.* **2024**, *57*, 6183–6209. [CrossRef]

12. Ohtsu, M.; Shiotani, T.; Shigeishi, M.; Kamada, T.; Yuyama, S.; Watanabe, T.; Suzuki, T.; van Mier, J.G.M.; Vogel, T.; Grosse, C.; et al. Recommendation of RILEM TC 212-ACD: Acoustic emission and related NDE techniques for crack detection and damage evaluation in concrete: Test method for classification of active cracks in concrete structures by acoustic emission. *Mater. Struct. Constr.* **2010**, *43*, 1187–1189.
13. Ohno, K.; Ohtsu, M. Crack classification in concrete based on acoustic emission. *Constr. Build. Mater.* **2010**, *24*, 2339–2346. [CrossRef]
14. Aggelis, D.G. Classification of cracking mode in concrete by acoustic emission parameters. *Mech. Res. Commun.* **2011**, *38*, 153–157. [CrossRef]
15. Zhang, Z.-H.; Deng, J.-H. A new method for determining the crack classification criterion in acoustic emission parameter analysis. *Int. J. Rock Mech. Min. Sci.* **2020**, *130*, 104323. [CrossRef]
16. Dong, L.; Zhang, Y.; Bi, S.; Ma, J.; Yan, Y.; Cao, H. Uncertainty investigation for the classification of rock micro-fracture types using acoustic emission parameters. *Int. J. Rock Mech. Min. Sci.* **2023**, *162*, 105292. [CrossRef]
17. Das, A.K.; Suthar, D.; Leung, C.K. Machine learning based crack mode classification from unlabeled acoustic emission waveform features. *Cem. Concr. Res.* **2019**, *121*, 42–57. [CrossRef]
18. Ju, S.; Li, D.; Jia, J. Machine-learning-based methods for crack classification using acoustic emission technique. *Mech. Syst. Signal Process.* **2022**, *178*, 109253. [CrossRef]
19. Qin, Z.; Chen, X.; Yan, Y.; Liu, Z.; Wang, Z. A SVM-based method for identifying fracture modes of rock using WVD spectrogram features of AE signals. *Front. Earth Sci.* **2023**, *11*, 1206269. [CrossRef]
20. Wang, J.; Liang, P.; Zhang, Y.; Yao, X.; Yu, G.; Han, Q. Real-time identification of acoustic emission signals of rock tension-shear fracture based on machine learning and study on precursory characteristics. *Mech. Syst. Signal Process.* **2025**, *230*, 112665. [CrossRef]
21. Torrence, C.; Compo, G.P. A Practical Guide to Wavelet Analysis. *Bull. Am. Meteorol. Soc.* **1998**, *79*, 61–78. [CrossRef]
22. Sun, L.; Lin, L.; Yao, X.; Zhang, Y.; Tao, Z.; Ling, P. Real-Time Recognition Method for Key Signals of Rock Fracture Acoustic Emissions Based on Deep Learning. *Sensors* **2023**, *23*, 8513. [CrossRef] [PubMed]
23. Ren, F.; Zhu, C.; Yuan, Z.; Karakus, M.; Tang, S.; He, M. Recognition of shear and tension signals based on acoustic emission parameters and waveform using machine learning methods. *Int. J. Rock Mech. Min. Sci.* **2023**, *171*, 105578. [CrossRef]
24. Sikdar, S.; Liu, D.; Kundu, A. Acoustic emission data based deep learning approach for classification and detection of damage-sources in a composite panel. *Compos. Part B Eng.* **2022**, *228*, 109450. [CrossRef]
25. Li, D.; Wang, Y.; Yan, W.-J.; Ren, W.-X. Acoustic emission wave classification for rail crack monitoring based on synchrosqueezed wavelet transform and multi-branch convolutional neural network. *Struct. Health Monit.* **2021**, *20*, 1563–1582. [CrossRef]
26. Song, Z.; Huang, J.; Deng, B.; Li, M.; Li, Q.; Liang, Q.; Zhang, C. Influence of acoustic emission sequence length on intelligent identification accuracy of 3-D loaded rock's fracture stage. *Eng. Fail. Anal.* **2024**, *162*, 108389. [CrossRef]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 770–778.
28. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 4700–4708. Available online: https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html (accessed on 10 October 2026).
29. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
30. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018*; IEEE: New York, NY, USA, 2018; pp. 4510–4520.
31. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljacic, M.; Hou, T.Y.; Tegmark, M. KAN: Kolmogorov-Arnold Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025. Available online: https://openreview.net/forum?id=Ozo7qJ5vZi (accessed on 10 October 2026).
32. Li, C.; Liu, X.; Li, W.; Wang, C.; Liu, H.; Liu, Y.; Chen, Z.; Yuan, Y. U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 4652–4660. [CrossRef]
33. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752.
34. Dao, T.; Gu, A. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research, Vienna, Austria, 21–27 July 2024*; PMLR: Cambridge, MA, USA, 2024; pp. 10041–10071. Available online: https://proceedings.mlr.press/v235/dao24a.html (accessed on 10 October 2026).
35. Bai, R.; Xu, Q.; Meng, Z.; Cao, L.; Xing, K.; Fan, F. Rolling Bearing Fault Diagnosis Based on Multi-Channel Convolution Neural Network and Multi-Scale clipping Fusion Data Augmentation. *Measurement* **2021**, *184*, 109885. [CrossRef]

36. Guo, F.; Li, W.; Jiang, P.; Chen, F.; Liu, Y. Deep Learning Approach for Damage Classification Based on Acoustic Emission Data in Composite Materials. *Materials* **2022**, *15*, 4270. [CrossRef] [PubMed]
37. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*; IEEE: New York, NY, USA, 2017; pp. 618–626.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.